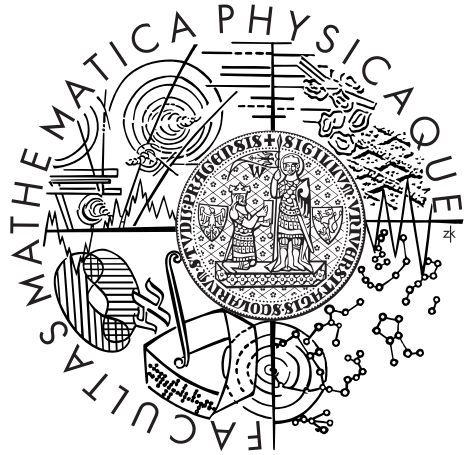
Karlova univerzita v Praze
Matematicko-fyzikální fakulta

Diplomová práce



Jozef Juríček

Odhad polohy nulových bodů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí: **Mgr. Zdeněk Hlávka, Ph.D.**

Studijní program: **Matematická statistika**

Studijní obor: **Pravděpodobnost, matematická statistika a ekonometrie**

Poděkování

Děkuji vedoucímu mé diplomové práce Mgr. Zdeňku Hlávkovi, Ph.D. za zadání zajímavého tématu diplomové práce a také za cenné rady, připomínky a návrhy při psaní práce.

Čestné prohlášení

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 20. dubna 2006 Jozef Juríček

Obsah

1 Parametrická regresia **5**
1.1 Odhad nulového bodu v parametrických modeloch 5
1.2 Lineárny model 6
1.2.1 Normalita 7
1.2.2 Konfidenčné množiny pre nulové body 9
1.3 Nelineárna regresia 13

2 Jadrová regresia **17**
2.1 Jadrové odhady v prípade *i.i.d.* 17
2.1.1 Odhady nulových bodov 20
2.1.2 Limitné rozdelenia 23
2.1.3 $\widehat{\sigma^2}$ 28
2.1.4 Jadrové funkcie 28
2.1.5 Vyhladzovací parameter 30
2.2 *Bootstrap* a konfidenčné množiny jadrových odhadov 34
2.3 Jadrové odhady s korelovanými chybami 37
2.3.1 Vyhladzovací parameter 38
2.3.2 $\widehat{\mathsf{cov}}$ 41
2.3.3 Známa variančná matica 42
2.3.4 Intervalové odhady nulových bodov 42
2.4 Aplikácia jadrových odhadov na reálne dáta 44

3 Prostredie R **48**
3.1 Príklad 1.21 48
3.2 Príklad 1.22 48
3.3 Príklad 1.25 48
3.4 Príklady 2.12, 2.20, tabuľky 1 a 2, kapitola 2.4 50

Název práce: Odhad polohy nulových bodů

Autor: Jozef Juríček

Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Zdeněk Hlávka, Ph.D.

e-mail vedoucího: hlavka@karlin.mff.cuni.cz

Abstrakt: Práce se zabývá odhady polohy nulových bodů regresní funkce a jejích derivací, a to jak postupy parametrické, tak neparametické regrese.

První část se věnuje parametrické regresi - lineárnímu i nelineárnímu modelu. Odhady polohy nulových bodů jsou pak založeny na odhadech parametrů.

Tématem druhé části je neparametrická regrese, v tomto případě jde o jádrové odhady navrženy Gasserem a Müllerem. Popisuje zejména limitní rozdělení odhadů, volbu vyhlazovacího parametru a jádrové funkce.

V obou částech jsou konstruovány intervalové odhady polohy nulových bodů regresní funkce a jejích derivací. Obě dvě části se věnují modelům s nezávislými, ale také s korelovanými chybami.

Práce nabízí i příklady k jednotlivým tématům, které jsou počítány v prostředí R a také některé zdrojové kódy funkcí nezbytných k výpočtům.

Klíčová slova: nulové body, neparametrická regrese, jádrová regrese, derivace, extrém.

Title: Estimation of Location of Zeros

Author: Jozef Juríček

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Zdeněk Hlávka, Ph.D.

Supervisor's e-mail address: hlavka@karlin.mff.cuni.cz

Abstract: The main interest of this master thesis is the estimation of location of zeros of the regression function and its derivatives by the parametric and nonparametric method.

The first section includes either linear and nonlinear regression model of the parametric methods. The estimators are then based on the estimates of parameters.

The second part includes nonparametric regression model - kernel estimators of the regression function and its derivatives investigated by Gasser and Müller. Especially, the limit distributions of the estimators of zeros and the choice of smoothing parameter and kernel function are studied.

Confidence bands for zeros of regression function and its derivatives are constructed in both sections. Models are studied with independent as well as correlated errors.

This master thesis offers examples to particular sections that are computed with software R and also sources of some programmed functions.

Keywords: zeros, nonparamteric regression, kernel regression, derivative, extremum.

Úvod

Témou práce je odhad polohy nulových bodov[1] regresnej fukcie a jej derivácií. V texte sa budeme zaoberať regresným modelom s jednou nezávisle premennou.

Analyzujeme dáta $(x_i, Y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$ a predpokladáme závislosť:

$$Y_i = f(x_i) + e_i, \qquad i = 1, \ldots, n,$$

kde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ je náhodný vektor známych "odpovedí" na vektor známych hodnôt vysvetľujúcich premenných $(x_1, \ldots, x_n)^T$, $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ je náhodný vektor chýb splňujúci $\mathsf{E}\boldsymbol{e} = \boldsymbol{0}$, $\mathsf{var}\ \boldsymbol{e}$ je konečná a $f\colon \mathbb{R} \to \mathbb{R}$ je reálna funkcia reálnej premennej.

Postupne budeme pridávať niektoré ďalšie predpoklady:

- na funkciu f (parametrický predpis, plná stĺpcová hodnosť regresnej matice, linearita v parametroch, spojitosť, diferencovateľnosť, atď.);
- na vektor chýb $\boldsymbol{e}$ (normalita, regularita variančnej matice, nekorelovanosť, nezávislosť, atď.);
- na hodnoty regresorov x_i (reštrikcia, ekvidistancia, atď.).

Budeme musieť byť opatrní najmä pri zavádzaní odhadu polohy nulového bodu, aby z teoretického hľadiska odhadom (teda borelovsky merateľnou funkciou náhodného vektoru $\boldsymbol{Y}$) vôbec bol.

Odhady polohy nulových bodov ζ_0 regresnej funkcie f, teda bodov, pre ktoré platí $f(\zeta_0) = 0$, budeme konštruovať najmä ako nulové body $\widehat{\zeta_0} = \zeta_{n,0}$ odhadu $\hat{f}$ funkcie f a pre tieto sa budeme snažiť odvodiť nejaké vlastnosti, prípadne aj nájsť konfidenčné množiny. Podobne pre odhad $\zeta_{n,\nu}$ nulového bodu ζ_ν ν-tej derivácie regresnej funkcie.

V parametrickom modeli budeme konfidenčné množiny odvodzovať predovšetkým z konfidenčných množín pre parametre regresnej funkcie.

V neparametrickom modeli jadrovej regresie budeme hľadať limitné vety, na základe ktorých ukážeme asymptotické vlastnosti našich odhadov a príslušné približné konfidenčné množiny. Zmienime sa aj o modernejšej *resampling* metóde (*bootstrap*) nájdenia konfidenčných množín.

K jednotlivým témam práce budú pridávané ilustračné príklady. Výsledky príkladov sú počítané a obrázky k príkladom sú generované softvérom R. V sekcii 3 sú uvedené podstatné časti zdrojového kódu k naprogramovaným funkciám a tiež niektoré postupy pri počítaní pomocou týchto funkcií v prostredí R.

[1] V celom texte sa budú stotožňovať pojmy "poloha nulového bodu" a "nulový bod".

1 Parametrická regresia

Kapitola pojednáva o odhadoch nulových bodov regresnej funkcie a jej derivácií na základe parametrického regresného modelu, nulové body bude odhadovať, pochopiteľne, na základe odhadov hodnôt parametra. Teoretické základy sú prevzaté najmä z [Zv04].

Aby sme mohli nulové body odhadovať v čo najväčšej triede funkcií a v čo najväčšom počte, budeme musieť postupovať z formálneho hľadiska opatrne.

Pre celú kapitolu vo všeobecnosti predpokladáme model:

$$Y_i = f(x_i, \boldsymbol{\theta}) + e_i; \quad i = 1, \ldots, n; \tag{1.1}$$

kde regresná funkcia $f(x, \boldsymbol{\theta})$ je spojitá v $x \in \mathbb{R}$ aj v $\boldsymbol{\theta} \in \mathbb{R}^k$ a známa až na $\boldsymbol{\theta}$, neznámy parameter modelu, ktorý budeme odhadovať a $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ je náhodný vektor s $\mathsf{E}\boldsymbol{e} = \mathbf{0}$; $\mathsf{var}\ \boldsymbol{e}$ konečná.

1.1 Odhad nulového bodu v parametrických modeloch

V celej kapitole 1.1 budeme formálne vychádzať z nasledujúcich definícií.

Definícia 1.1 (Pseudoinverzná funkcia nulového bodu). Ľubovoľnú borelovsky merateľnú funkciu $f^- : (\mathbb{R}^k, \mathbb{B}^k) \to (\overline{\mathbb{R}}, \overline{\mathbb{B}})$ takú[2][3], že $\forall_{\boldsymbol{\theta} \in \mathbb{R}^k} f^-(\boldsymbol{\theta}) = \zeta_{\boldsymbol{\theta}}$, pričom $f(\zeta_{\boldsymbol{\theta}}) = 0$, dodefinujúc $f(-\infty) = 0, f(\infty) = \infty$ budeme nazývať *pseudoinverzná funkcia nulového bodu funkcií* f. Množinu $\mathcal{F}_0^- = \{\boldsymbol{\theta} \in \mathbb{R}^k : f^-(\boldsymbol{\theta}) = -\infty\}$ môžeme nazvať *množina nenulovosti funkcií* f.

Poznámka 1.2. Existencia pseudoinverznej funkcie nulového bodu je zaručená predpokladom spojitosti funkcie f a axiómom výberu, ktorý nám v prípade viacerých možností polohy nulového bodu zaručuje "schopnosť výberu".

Reálnu priamku sme si rozšírili pre prípad, že náš odhad regresnej funkcie nebude mať nulový bod; len preto, aby sme mohli "prejsť" všetky parametre a pre tie, pre ktoré f nebude mať nulový bod, budeme považovať polohu nulového bodu za nevlastný bod $-\infty$.

Poznámka 1.3 *(Súvis s pseudoinverznou maticou).* Ak uvážime viacrozmernú analógiu def. 1.1, potom pseudoinverznou funkciou nulového bodu (bodu jadra) homomorfizmov $\boldsymbol{\Lambda}_{n\times m}\boldsymbol{x}_{m\times 1} - \boldsymbol{\beta}_{n\times 1}$, môže byť ľubovoľné zobrazenie $\boldsymbol{\Lambda}^-\boldsymbol{\beta}$, pre $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times m}$ a $\boldsymbol{\beta} \in \mathbb{R}^n$ parametre také, že $h(\boldsymbol{\Lambda}) = h(\boldsymbol{\Lambda}|\boldsymbol{\beta})$ a $(-\infty, \ldots, -\infty)^T{}_{m\times 1}$ pre nekonzistentné sústavy. Výber pseudoinverznej matice $\boldsymbol{\Lambda}^- \in \mathbb{R}^{m\times n}$ je skoro vždy na nás a na axióme výberu.

[2] $\mathbb{B}^k$ ... borelovská σ-algebra priestoru $\mathbb{R}^k$.

[3] $\overline{\mathbb{B}}$ ... borelovská σ-algebra priestoru $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Príklad 1.4 (Pseudoinverzná funkcia nulového bodu priamok). Nech $f(x) = \alpha + \beta x$. Potom zaveďme

$$f^-(\alpha,\beta) = \begin{cases} -\frac{\alpha}{\beta} & \text{pre } \alpha \in \mathbb{R} \setminus \{0\}, \beta \in \mathbb{R} \setminus \{0\} \\ -\infty & \text{pre } \alpha \in \mathbb{R} \setminus \{0\}, \beta = 0 \\ \text{vyberieme si} & \text{pre } (\alpha,\beta) = (0,0) \end{cases} . \tag{1.2}$$

Práve zavedený formalizmus zaručuje korektnosť definície 1.5.

Definícia 1.5 (Odhad polohy nulového bodu). Odhadom polohy nulového bodu v parametrických regresných modeloch s jednou nezávisle premennou bude zobecnená náhodná veličina[4] $f^-(\boldsymbol{t})$, kde $\boldsymbol{t}$ je odhadom parametra $\boldsymbol{\theta}$.

1.2 Lineárny model

V tejto kapitole najprv v krátkosti uvedieme niektoré známe tvrdenia o odhadoch parametrov v lineárnom modeli s plnou hodnosťou. V závere kapitoly odvodíme vetu o odhade nulového bodu v lineárnom modeli, a to pre dosť obecnú triedu funkcií.

Predpokladáme, že

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.3}$$

kde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T \in \mathbb{R}^k$ je vektor neznámych parametrov, $\boldsymbol{e} = (e_1, \ldots, e_n)$ je náhodný vektor a $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ známa matica, $h(\boldsymbol{X}) = k$; i-ty riadok matice $\boldsymbol{X}$ je (transponovaný) k-rozmerný vektor $\boldsymbol{g}(x_i) = (g_1(x_i), \ldots, g_k(x_i))^T$.

$$\boldsymbol{X} = \begin{pmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_k(x_1) \\ g_1(x_2) & g_2(x_2) & \ldots & g_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(x_n) & g_2(x_n) & \ldots & g_k(x_n) \end{pmatrix} \tag{1.4}$$

Takto dosiahneme súlad so značením v úvode, f bude lineárnou kombináciou funkcií $g_j, j = 1, \ldots, k$: $f = \sum_{j=1}^k g_j\beta_j = \boldsymbol{g}^T\boldsymbol{\beta} = \boldsymbol{\beta}^T\boldsymbol{g}$. Pod linearitou modelu v tomto prípade rozumieme linearitu funkcie f v $\boldsymbol{\beta}$. Model je potom možné vyjadriť v tvare:

$$Y_i = \boldsymbol{\beta}^T\boldsymbol{g}(x_i) + e_i, \qquad i = 1, \ldots n. \tag{1.5}$$

Veta 1.6 (Odhad strednej hodnoty a jej vlastnosti I). *Nech platí model* (1.5). *Ak* $\mathsf{E}\boldsymbol{e} = \boldsymbol{0}$ *a* $\mathsf{var}\ \boldsymbol{e} = \sigma^2\boldsymbol{I}$, *potom najlepším nestranným lineárnym odhadom parametra* $\boldsymbol{\beta}^T\boldsymbol{g}(x) =: \mathsf{E}Y(x)$ *je*

$$\hat{f}(x) := \widehat{Y(x)} := \boldsymbol{b}^T\boldsymbol{g}(x), \quad kde \quad \boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{1.6}$$

$$a \quad \mathsf{var}\ \widehat{Y(x)} = \sigma^2\left(\boldsymbol{g}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{g}(x)\right). \tag{1.7}$$

[4] $f^-(\boldsymbol{t})$ s hodnotami v $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$.

Nestranným odhadom parametra σ^2 *je*

$$s^2 = \frac{1}{n-k}\boldsymbol{Y}^T\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]\boldsymbol{Y}. \tag{1.8}$$

Dôkaz. [Zv04]. □

Veta 1.7 (Odhad strednej hodnoty a jej vlastnosti W). *Nech platí model* (1.5). *Ak* $\mathsf{E}\boldsymbol{e} = \boldsymbol{0}$ *a* $\mathsf{var}\ \boldsymbol{e} = \sigma^2\boldsymbol{W}^{-1}$, $\boldsymbol{W} > 0$, *potom najlepším nestranným lineárnym odhadom parametra* $\mathsf{E}Y(x)$ *je*

$$\hat{f}(x) := \widehat{Y(x)} := \boldsymbol{b}^T\boldsymbol{g}(x), \quad kde \quad \boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \tag{1.9}$$

$$a \quad \mathsf{var}\,\widehat{Y(x)} = \sigma^2\left(\boldsymbol{g}^T(x)(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{g}(x)\right). \tag{1.10}$$

Nestranným odhadom parametra σ^2 *je*

$$s^2_{\boldsymbol{W}} = \frac{1}{n-k}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\right)^T\boldsymbol{W}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\right). \tag{1.11}$$

Dôkaz. Napríklad prechodom k modelu $\boldsymbol{C}\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{e}$, kde $\boldsymbol{C} = \boldsymbol{W}^{\frac{1}{2}}$, ako uvádza [Zv04], potom z predchádzajúcej vety. □

Poznámka 1.8 *(Derivácie).* Ako odhad ν-tej derivácie sa ponúka ν-krát zderivovať odhad $\widehat{Y(x)}$. Je zrejmé, že jeho vlastnosti budú "podobné". Vyplýva to z toho, že v našom modeli je funkcia f lineárnou kombináciou funkcií g_j a z vlastnosti linearity derivácie. Musíme však, prirodzene, pridať ešte predpoklad príslušnej hladkosti na funkcie g_j, resp. na f, teda, že $g_j \in \mathcal{C}^\nu$, $j = 1, \ldots, k$.

1.2.1 Normalita

Pridajme predpoklad normality na rozdelenie chýb $\boldsymbol{e}$, pričom zachováme $\mathsf{E}\boldsymbol{e}$, $\mathsf{var}\ \boldsymbol{e}$ z viet 1.6, resp. 1.7. Poznáme už nestranné odhady parametrov a týmto aj ich rozdelenie. Môžeme teda sformulovať nasledujúce tvrdenia.

Veta 1.9 (Interval spoľahlivosti I). *Nech sú splnené predpoklady vety 1.6. Nech* $\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. *Potom platí:*

$$\frac{(n-k)s^2}{\sigma^2} \quad \sim \quad \chi^2_{n-k} \tag{1.12}$$

$$a \quad \widehat{Y(x)} \quad \pm \quad t_{n-k}(\alpha)s\sqrt{\left(\boldsymbol{g}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{g}(x)\right)} \tag{1.13}$$

sú[5] *krajné body* $100(1-\alpha)\%$*-ného intervalu spoľahlivosti pre funkčnú hodnotu regresnej funkcie v každom predom pevne zvolenom bode* x.

[5] Kritické hodnoty definované v [An02]... pre $T \sim \mathsf{t}_{n-k}$: $\mathsf{P}[|T| \geqq t_{n-k}(\alpha)] = \alpha$.

Dôkaz. Aplikáciou vzťahu medzi normálnym rozdelením a jeho kvadratickými formami a vety 1.6. □

Veta 1.10 (Interval spoľahlivosti W). *Nech sú splnené predpoklady vety 1.7. Nech* $\boldsymbol{e} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{W}^{-1}), \boldsymbol{W} > 0$. *Potom platí:*

$$\frac{(n-k)s_W^2}{\sigma^2} \sim \chi^2_{n-k} \tag{1.14}$$

$$a \quad \widehat{Y(x)} \pm t_{n-k}(\alpha) s_{\boldsymbol{W}} \sqrt{\left(\boldsymbol{g}^T(x)(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{g}(x)\right)} \tag{1.15}$$

sú krajné body $100(1-\alpha)\%$*-ného intervalu spoľahlivosti pre funkčnú hodnotu regresnej funkcie v každom predom pevne zvolenom bode* x.

Dôkaz. Použitím viet 1.9 a 1.7. □

Poznámka 1.11. Ak nehľadáme interval spoľahlivosti pre funkčnú hodnotu, ale predikčný interval pre nezávislé budúce pozorovanie $Y(x) = \boldsymbol{\beta}^T \boldsymbol{g}(x) + e$, potom vzroce pre krajné body týchto intervalov budú analogické ako vo vetách 1.9, resp. 1.10, akurát zátvorka pod odmocninou sa zväčší o $\frac{\mathsf{var}\, e}{\sigma^2}$.

Skutočnú hodnotu regresnej funkcie v predom pevne danom bode x pokrýva interval spoľahlivosti s pravdepodobnosťou $1-\alpha$. Ak budeme vytvárať takýto interal postupne pre každé x (jeho hranice chápať ako funkciu premennej x), dostaneme *pás spoľahlivosti okolo regresnej krivky.* V kap. 1.2.2 budeme potrebovať skúmať aj množinu, ktorá pokrýva celú regresnú krivku súčasne s pravdepodobnosťou $1-\alpha$. Takúto množinu nazývame *pás spoľahlivosti pre regresnú krivku.*

Veta 1.12 (Pás spoľahlivosti pre regresnú krivku). *Nech sú splnené predpoklady vety 1.9. Potom množina*

$$\mathcal{K} = \left\{ (x,y) \in \mathbb{R}^2 : |y - \boldsymbol{b}^T \boldsymbol{g}(x)| \leqq s \cdot \sqrt{\boldsymbol{g}^T(x)(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{g}(x) \cdot k \cdot F_{k,n-k}(\alpha)} \right\} \tag{1.16}$$

pokrýva[6] *celú regresnú krivku (súčasne) s pravdepodobnosťou rovnou najmenej* $1-\alpha$.

Dôkaz. Nájde sa v [Zv04]. Vychádza sa z konfidenčnej množiny pre parameter $\boldsymbol{\beta}$:

$$\mathcal{K}_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} \in \mathbb{R}^k : (\boldsymbol{\beta} - \boldsymbol{b})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \boldsymbol{b}) \leqq s^2 k \cdot F_{k,n-k}\}. \tag{1.17}$$

□

Poznámka 1.13 *(Korelované chyby).* Podobne, ak sú splnené predpoklady vety 1.10, platí:

$$\mathcal{K} = \left\{ (x,y) \in \mathbb{R}^2 : |y - \boldsymbol{b}^T \boldsymbol{g}(x)| \leqq s_{\boldsymbol{W}} \cdot \sqrt{\boldsymbol{g}^T(x)(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{g}(x) \cdot k \cdot F_{k,n-k}(\alpha)} \right\}. \tag{1.18}$$

[6] Kritické hodnoty definované v [An02]... pre $Z \sim \mathsf{F}_{k,n-k}$: $\mathsf{P}[Z \geqq F_{k,n-k}(\alpha)] = \alpha$.

1.2.2 Konfidenčné množiny pre nulové body

Pre pohodlné odvodenie konfidenčných množín pre nulové body budeme potrebovať predpoklady 1.14.

Predpoklady 1.14 (Lineárna nezávislosť). Pre ďalšie zjednodušenie predpokladajme, že $g_1 \equiv 1$, $g_2, \ldots, g_k$ sú spojité a funkcie $1, g_2, \ldots, g_k$ sú lineárne nezávislé na každom *otvorenom nedegenerovanom intervale*[7][8].

Poznámka 1.15. Silné predpoklady 1.14 nám pohodlne zaručili, že funckia f bude mať pre každý parameter $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^k$ spočetnú množinu nulových bodov.

Hľadajme teda konfidenčnú množinu pre polohu nulového bodu. Mohli by sme uvažovať množinu

$$\mathcal{K}_0 = \{x \in \mathbb{R} : (x, 0) \in \mathcal{K}\} \tag{1.19}$$

Táto konfidenčná množina však zahŕňa odhady všetkých nulových bodov s pravdepodobnosťou nie menšou ako $1 - \alpha$. Ak by sme chceli nájsť konfidenčnú množinu nejakého konkrétneho nulového bodu[9], môžu sa nám konfidenčné množiny viacerých nulových bodov "prekryť" - v tomto prípade môžeme uvažovať množinu $f^-(\mathcal{K}_{\boldsymbol{\beta}})$. Uvedomme si, že hodnoty parametrov $\beta_1, \ldots, \beta_k$ sú polohou nulového bodu $f^-(\boldsymbol{\beta})$ v určitom zmysle viazané. V takomto prípade nám k určeniu konfidenčných množín pre jednotlivé nulové body poslúži veta 1.16.

Veta 1.16 (Konfidenčná množina pre odhad polohy nulového bodu). *Nech sú splnené predpoklady 1.14. Nech* $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$*, potom množina*

$$\mathcal{K}_0^* = \left\{ x \in \mathbb{R} : |\boldsymbol{b}^T\boldsymbol{g}(x)| \leqq s \cdot t_{n-k}(\alpha)\sqrt{\boldsymbol{g}^T(x)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{g}(x)} \right\}. \tag{1.20}$$

je konfidenčnou množinou pre odhad nulového bodu $f^-(\boldsymbol{b})$ *s pravdepodobnosťou nie menšou ako* $1 - \alpha$

Dôkaz. Množina je založená na testovaní hypotézy $\zeta = \zeta_0$. Tvrdenie plynie podľa Scheffého vety (napr. v [An02], veta 10.2, str. 206), ak si uvedomíme, že pre pevné ζ_0 je funkcia $f = \boldsymbol{\beta}^T\boldsymbol{g}(\zeta_0)$ homomorfizmom (v $\boldsymbol{\beta}$) a ak si uvedomíme že

$$\dim\big(\mathrm{Ker}(\boldsymbol{\gamma}^T\boldsymbol{g}(x))\big) = k - 1 \quad \text{pre } \lambda \text{ - skoro všetky}[10] \text{ možné nulové body } x \in \mathbb{R}. \tag{1.21}$$

Potom podľa vety o jadre a obraze homomorfizmu

$$\dim\big(\mathrm{Im}(\boldsymbol{\gamma}^T\boldsymbol{g}(x))\big) = k - (k-1) = 1. \tag{1.22}$$

[7] *Otvorený nedegenerovaný interval* $\ldots \mathcal{I} = (a, b)$, pričom $-\infty \leqq a < b \leqq \infty$.
[8] *Lineárna nezávislosť na* $\mathcal{I}$ $\ldots$ na $\mathcal{I}$ platí: $[\boldsymbol{\gamma}^T\boldsymbol{g}(\cdot) \equiv 0] \Leftrightarrow [\boldsymbol{\gamma} \equiv \mathbf{0}], \boldsymbol{\gamma} \in \mathbb{R}^k$.
[9] "Konkrétny" nulový bod je daný predpisom $f^-(\boldsymbol{\theta})$.
[10] $\lambda \ldots$ Lebesgueova miera na $\mathbb{R}$.

Hypotézu zamietneme, ak $\boldsymbol{\beta} \in \mathrm{Im}(\boldsymbol{\gamma}^T \boldsymbol{g}(x))$. Obraz homomorfizmu je vektorový priestor. Preto prechádzame parametrom $\boldsymbol{\beta}$ iba podpriestor dimenzie 1. Uvedomme si ešte, že $t_{n-k}(\alpha) = \sqrt{1 \cdot F_{1,n-k}(\alpha)}$.

V patologických prípadoch sa dostaneme až do bodu $-\infty$. □

Poznámka 1.17. Vzťah (1.21) vlastne hovorí, že jedna súradnica pre pevný nulový bod sa dá dopočítať z ostatných súradníc, čo je očakávateľné, ale hlavne, že pre skoro všetky možné nulové body potrebujeme všetky ostatné súradnice[11]. Skutočnosť, že sa jedná o homomorfizmus (v $\boldsymbol{\beta}$), napovedá, že v rámci nelineárnej regresie (kap. 1.3) tento princíp nebude využiteľný (nelinearita v $\boldsymbol{\beta}$), pokiaľ si nevsadíme na lineárne priblíženie funkcie f (v $\boldsymbol{\beta}$).

Poznámka 1.18. Konfidenčná množina nulového bodu je vlastne množina tých bodov x, pre ktoré bod 0 "padne" do intervalu spoľahlivosti $\boldsymbol{I}$ (veta 1.9) pre $f(x)$.

Poznámka 1.19. Ak by sme chceli odhadovať nulové body na nejakej podmnožine $\mathbb{R}$, mohlo by to implicitne "vyvolať" aj obmedzenia na parameter $\boldsymbol{\beta}$. Ak teda vieme predom polohu nulového bodu lokalizovať na nejakú podmnožinu, môžeme zúžiť aj množinu prípustných parametrov.

Prípad lineárnych obmedzení na parametre a predom danej hodnoty parametra sú obecne popísané v [Zv04].

Poznámka 1.20. Analogické tvrdenie možno vysloviť aj pre model s korelovanými chybami.

Príklad 1.21 (Priamka). Skúmame dáta $(x_1, Y_1), \ldots (x_{11}, Y_{11})$ - našou úlohou je odhadnúť nulový bod a nájsť preň konfidenčnú množinu, pričom uvažujeme model:

$$Y_i = \beta_0 + \beta_1 x_i + e_i; \quad e_i \sim i.i.d.\ \mathsf{N}(0, \sigma^2); \quad i = 1, \ldots, 11.$$

V skutočnosti $f(x) = 1.6x - 18$, $\zeta_0 = 11.25$, $\sigma = 1$.

Obr. 1 znázorňuje jednotlivé pozorovania, náš odhad priamky, bodový a intervalový odhad nulového bodu založený na vete 1.16.

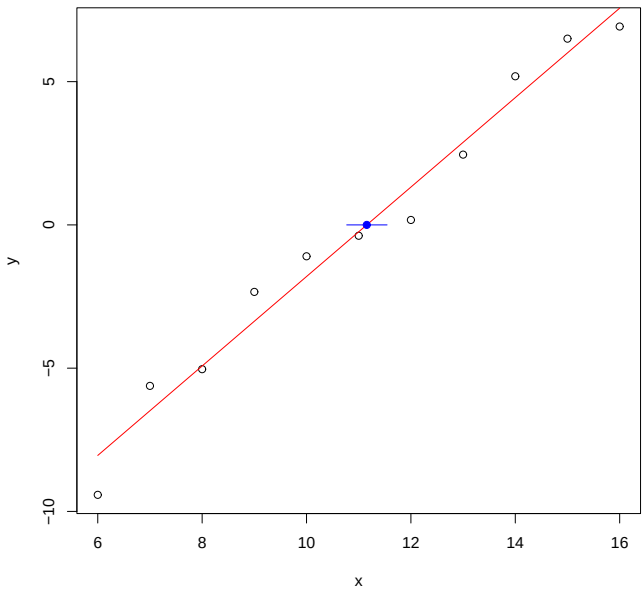
Krajné body intervalu na zálade vety 1.16, resp. vzťahu (1.20) vyjdú:

`x.left=10.76978` `x.est=11.15433` `x.right=11.54350` (obr. 1)

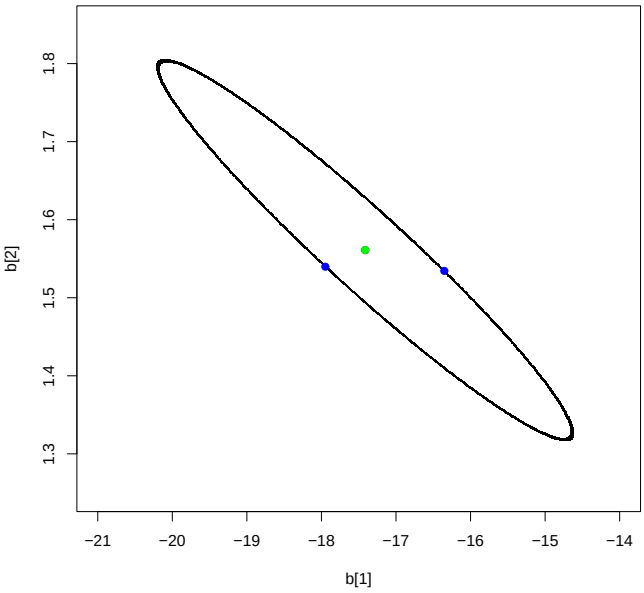
Na obr. 2 je znázornená konfidenčná elipsa pre parametre modelu (vzorec (1.17)), znázornené sú taktiež body (b_1, b_2), v ktorých je nulový bod funkcie $y = b_1 + b_2 x$ krajným bodom intervalu spoľahlivosti na základe (1.19). Interval vyjde o málo širší ako interval na obr. 1:

`x.left=10.65693` `x.est=11.15433` `x.right=11.65606`

[11] Toto tvrdenie nie je až také triviálne. Treba uvážiť "naraz" spojitosť funkcií g_j a ich lineárnu nezávislosť v zmysle predpokladov 1.14. Fakt, že na kraji konfidenčnej množiny bude práve x také, že dimenzia jadra príslušného homomorfizmu bude $k-1$, potom vyplýva z "nespojitosti F rozdelenia v stupňoch voľnosti".



Obr. 1: Bodový a 95% intervalový odhad nulového bodu priamky



Obr. 2: 95% konfidenčná elipsa pre parametre modelu s označeným stredom a parametre pre krajné body intervalu spoľahlivosti

Príklad 1.22 (Parabola). Skúmame dáta $(x_1, Y_1), \ldots (x_{21}, Y_{21})$ a predpokladáme model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i; \quad e_i \sim i.i.d.\ \mathsf{N}(0, \sigma^2); \quad i = 1, \ldots, 21.$$

Našou úlohou je odhadnúť "ľavý" nulový bod paraboly a nájsť preň konfidenčnú množinu so spoľahlivosťou 95%.

V skutočnosti $f(x) = 4x^2 + 9x - 9$, $\zeta_0 = -3$, $\sigma = 10$.

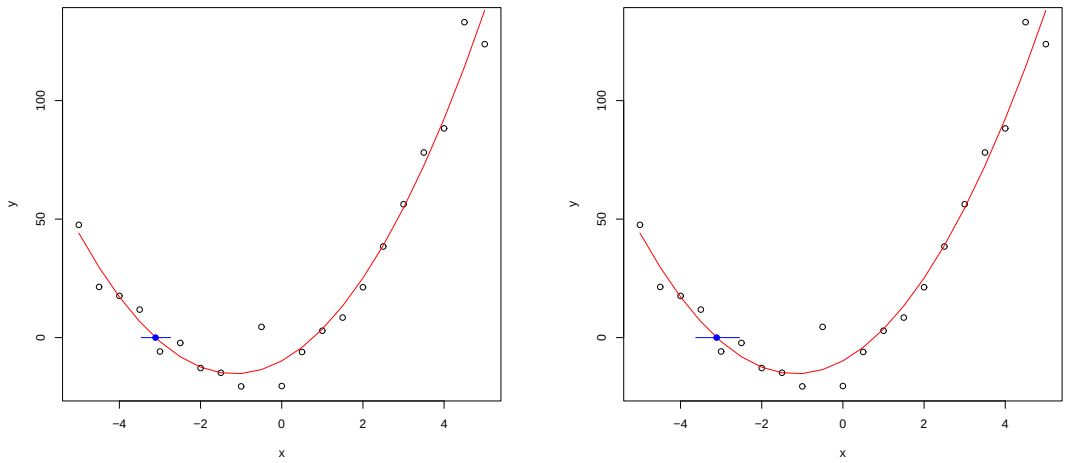
Podobne ako v príklade 1.21 dostaneme postupne 95% intervalové odhady. Na základe (1.20) máme (obr. 3 vľavo):

`x.left=-3.461100  x.est=-3.109863  x.right=-2.739240`

Zo vzorca (1.17) dostaneme (obr. 3 vpravo):

`x.left=-3.626406  x.est=-3.109863  x.right=-2.547842`

Obrázok 3 porovnáva 95% konfidenčné množiny pre nulový bod paraboly získané na základe vety 1.16 a ako $f^-(\mathcal{K}_{\boldsymbol{\beta}})$.



Obr. 3: Bodové a 95% intervalové odhady "ľavého" nulového bodu paraboly

Poznámka 1.23. Vyšetrovanie priebehu funkcie je v lineárnych modeloch "dostatočne" vierohodné, ak je "dostatočne" veľká hodnota koeficientu determinácie. Postupy, ktorými dosiahneme čo najprijateľnejší model, sú popísané v [Zv04].

1.3 Nelineárna regresia

V tejto kapitole sa budeme snažiť odhadovať parametre pre obecnejší tvar regresnej funkcie. Využívať budeme predovšetkým lineárne priblíženie z prvých dvoch členov Taylorovho rozvoja funkcie ("podľa parametra"), čím prevedieme úlohu na prípad kapitoly 1.2. Čerpáme z [Zv04]. Uvažujeme model:

$$Y_i = f(x_i, \boldsymbol{\theta}) + e_i, \qquad i = 1, \ldots, n. \tag{1.23}$$

Nech sú splnené ďalšie predpoklady:

(P1) $e_i \sim i.i.d.\ \mathsf{N}(0, \sigma^2)$;

(P2) $\boldsymbol{\theta} \in \Omega, \Omega \subset \mathbb{R}^k$ otvorená, konvexná;

(P3) $f \in \mathcal{C}^2(\Omega)$ pre všetky $x \in \mathcal{X}$;

(P4) matica $\boldsymbol{F}(\boldsymbol{\theta})$ prvých parciálnych derivácií regresnej funkcie f typu $n \times k$ daná vzťahom $\boldsymbol{F}(\boldsymbol{\theta}) = (f_j(x_i, \boldsymbol{\theta}))$, kde $f_j(x, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} f(x, \boldsymbol{\theta})$, má v okolí skutočnej hodnoty parametra $\boldsymbol{\theta}^*$ hodnosť k.

Odhad $\boldsymbol{t}$ parametra $\boldsymbol{\theta}$ metódou najmenších štvorcov dostaneme minimalizáciou reziduálneho súčtu štvorcov

$$\mathsf{RSS}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (Y_i - f(x_i, \boldsymbol{\theta}))^2, \boldsymbol{\theta} \in \Omega. \tag{1.24}$$

Ako odhad rozptylu použijeme

$$s^2 = \frac{\mathsf{RSS}(\boldsymbol{t})}{n - k}. \tag{1.25}$$

Odhad s^2 je asymptoticky ekvivalentný s odhadom rozptylu metódou maximálnej vierohodnosti $\frac{S(\boldsymbol{t})}{n}$.

Tento postup vedie k normálnej rovnici (položením parciálnych derivácií rovných nule):

$$\boldsymbol{F}(\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\theta})) = \boldsymbol{0}, \text{ kde } \boldsymbol{f}(\boldsymbol{\theta}) = (f(x_1, \boldsymbol{\theta}), \ldots, f(x_n, \boldsymbol{\theta})). \tag{1.26}$$

Pre $\boldsymbol{\theta}$ blízke $\boldsymbol{\theta}^*$ použijeme lineárnu aproximáciu

$$\boldsymbol{f}(\boldsymbol{\theta}) \doteq \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{F}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \tag{1.27}$$

Po dosadení aproximácie do normálnej rovnice a využitím normality dostávame aproximatívne rozdelenie $\boldsymbol{t}$:

$$\boldsymbol{t} \dot{\sim} \mathbf{N}(\boldsymbol{\theta}^*, \sigma^2(\boldsymbol{F}^{*T}\boldsymbol{F}^*)^{-1}), \text{ kde } \boldsymbol{F}^* = \boldsymbol{F}(\boldsymbol{\theta}^*). \tag{1.28}$$

Podobne pre reziduálny súčet štvorcov dostávame aproximáciu

$$\mathsf{RSS}(\boldsymbol{t}) \doteq \|(\boldsymbol{I} - \boldsymbol{F}^*(\boldsymbol{F}^{*T}\boldsymbol{F}^*)^{-1}\boldsymbol{F}^{*T})\boldsymbol{e}\|^2 \dot{\sim} \sigma^2\chi^2_{n-k}. \tag{1.29}$$

Asymptotická nezávislosť $\boldsymbol{t}$ a $\mathsf{RSS}(\boldsymbol{t})$ a konzistencia odhadu $\boldsymbol{t}$ nás vedie k ďalšej aproximácii

$$\frac{t_j - \theta_j^*}{s \cdot \sqrt{v_{jj}}} \dot{\sim} \mathsf{t}_{n-k} \ , \ j = 1, \ldots, k, \tag{1.30}$$

kde v_{jj} je diagonálny prvok matice $\boldsymbol{V} = (\boldsymbol{F}(\boldsymbol{t})^T\boldsymbol{F}(\boldsymbol{t}))^{-1}$.

Zo vzťahu (1.30) získame približný konfidenčný interval pre jednu zložku parametra θ_j:

$$(t_j - s \cdot \sqrt{v_{jj}}t_{n-k}(\alpha), t_j + s \cdot \sqrt{v_{jj}}t_{n-k}(\alpha)). \tag{1.31}$$

Z aproximácií (1.28) a (1.29) a z [An02], pozn 8.17, str. 178 potom dostávame konfidenčné množiny pre celý parameter $\boldsymbol{\theta}$ založené na testovaní hypotézy $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Waldov test dá:

$$\mathcal{K}^W_{\boldsymbol{\theta}} = \left\{\boldsymbol{\theta} \in \Omega; (\boldsymbol{\theta} - \boldsymbol{t})^T\boldsymbol{F}(\boldsymbol{t})^T\boldsymbol{F}(\boldsymbol{t})(\boldsymbol{\theta} - \boldsymbol{t}) < ks^2F_{k,n-k}(\alpha)\right\}. \tag{1.32}$$

Test pomerom vierohodností (vierohodnostnej funkcie v $\boldsymbol{t}$ a $\boldsymbol{\theta}$) dáva:

$$\mathcal{K}^{LR}_{\boldsymbol{\theta}} = \left\{\boldsymbol{\theta} \in \Omega; \mathsf{RSS}(\boldsymbol{\theta}) < \mathsf{RSS}(\boldsymbol{t})\left(1 + \frac{k}{n-k}F_{k,n-k}(\alpha)\right)\right\}. \tag{1.33}$$

Približné konfidenčné množiny pre polohu nulového bodu dopočítame z (1.32), resp. z (1.33):

$$\mathcal{K}^W_0 = \left\{x \in \mathbb{R}; f^-(\boldsymbol{\theta}) = x, \boldsymbol{\theta} \in \mathcal{K}^W_{\boldsymbol{\theta}}\right\}, \tag{1.34}$$

$$\mathcal{K}^{LR}_0 = \left\{x \in \mathbb{R}; f^-(\boldsymbol{\theta}) = x, \boldsymbol{\theta} \in \mathcal{K}^{LR}_{\boldsymbol{\theta}}\right\}. \tag{1.35}$$

Podobne môžeme konštruovať odhady nulových bodov derivácií f a ich približné konfidenčné množiny.

Ak napríklad vopred nepoznáme tvar regresnej funkcie a stanovujeme ho na základe"tvaru" dát, ale pritom vopred vieme, že máme odhadovať derivácie, odhady ich nulových bodov môžeme založiť aj na transformovaných dátach (pomocou "diskrétneho tvaru" Lagrangeovej vety). Takže označme:

$$x_i^{(0)} := x_i, \quad Y_i^{(0)} := Y_i, \quad i = 1, \ldots, n \tag{1.36}$$

$$x_i^{(\nu)} := \frac{1}{2}(x_{i+1}^{(\nu-1)} + x_i^{(\nu-1)}), \quad Y_i^{(\nu)} := \frac{Y_{i+1}^{(\nu-1)} - Y_i^{(\nu-1)}}{x_{i+1}^{(\nu-1)} - x_i^{(\nu-1)}}, \quad i = 1, \ldots n-\nu, \nu \geqq 1.$$

Potom prejdeme k modelu, ktorý môžeme symbolicky zapísať:

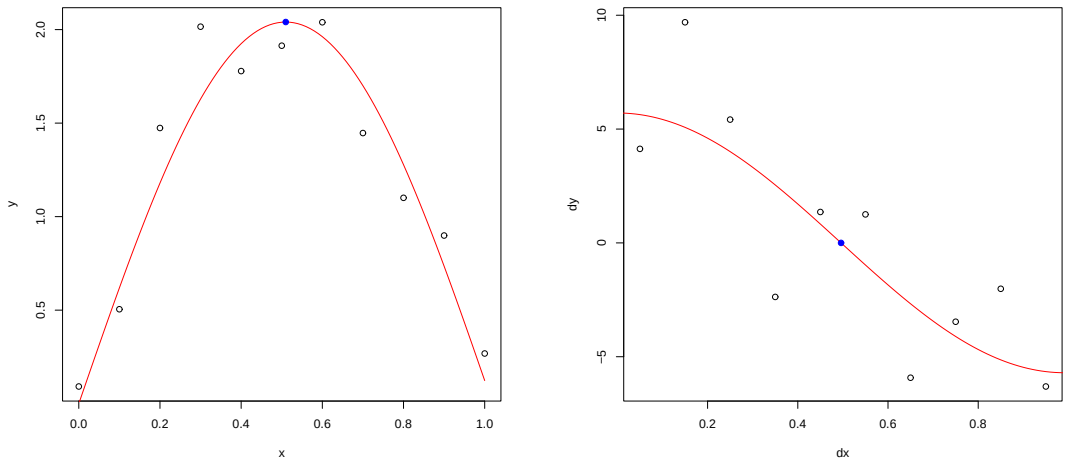
$$Y_i^{(\nu)} = f^{(\nu)}(x_i^{(\nu)}, \boldsymbol{\theta}) + e_i^{(\nu)}, \quad \text{pre } i = 1, \ldots, n-\nu \tag{1.37}$$

Poznámka 1.24. Uvedomme si, že sme transformáciami (1.36) mohli narušiť prípadnú nezávislosť alebo naopak, transformácia mohla znížiť závislosť pôvodných reziduí.

Príklad 1.25 (Porovanie odhadov polohy extrému a nulového bodu prvej derivácie). Skúmame dáta $(x_1, Y_1), \ldots (x_{11}, Y_{11})$ a predpokladáme model:

$$Y_i = a \cdot \sin(c \cdot \pi x_i) + e_i; \quad e_i \sim i.i.d\, \mathsf{N}(0, \sigma^2); \quad i = 1, \ldots, 11.$$

V skutočnosti $f(x) = 2\sin(\pi x)$, $x_{max} = 0.5$, $\sigma = 0.5$.



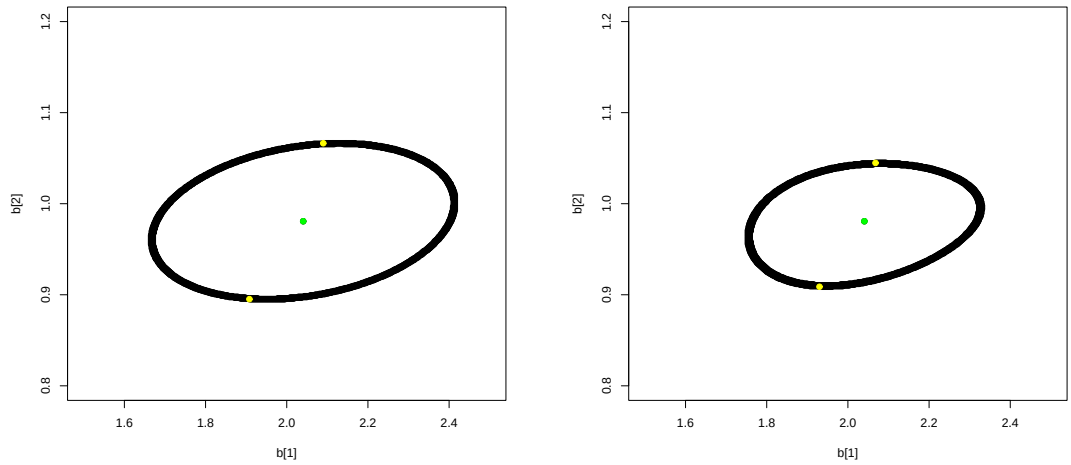
Obr. 4: Odhad polohy extrému (= 0.510) vs. nulového bodu 1. derivácie (= 0.496)

Na obr. 4 je znázornené porovnanie odhadov polohy extrému a polohy nulového bodu prvej derivácie regresnej funkcie na základe odhadu parametrov regresnej funkcie, resp. jej prvej derivácie. V pravej časti obr. 4 si tiež všimnime transformované dáta podľa (1.36).

Pre zaujímavosť na obr. 5 načrtneme konfidenčné množiny pre vektorový parameter založené na Waldovom teste (1.32) - v tomto prípade dostaneme elipsu, resp. teste pomerom vierohodností (1.33) - množina bude mať zložitejší tvar. Označíme tiež body, v ktorých sú dosiahnuté hranice približných intervalov spoľahlivosti (1.34) a (1.35).

Približné 95% intervalové odhady pre extrém založené na (1.34), resp. (1.35) vyjdú:

```
x.left=0.4688672    x.est=0.5098297    x.right=0.5585344
x.left=0.4785605    x.est=0.5098297    x.right=0.5501761
```



Obr. 5: 95% konfidenčné množiny pre parametre (a, c)

V celej kapitole sa až príliš spoliehame na vhodnosť lineárnej aproximácie (a, ako v celej parametrickej regresii, na samotný parametrický predpis funkcie f).

"Jemnejšie" aproximácie a problémy s tým spojené (napr. závislosť na parametrizácii) sú riešené v [Zv04].

2 Jadrová regresia

V nasledujúcich odstavcoch spomenieme modernejší a robustnejší prístup k regresii a ponúkneme takto ďalšie možnosti odhadovania nulových bodov regresnej funkcie, resp. jej derivácií. Jadrové odhady regresnej funkcie sú podtriedou tzv. neparametrických regresných odhadov. Hlavnou výhodou neparametrického prístupu v porovnaní s parametrickým je hlavne širšia trieda možných odhadov. Zaoberať sa budeme jednorozmerným prípadom. Teoretickým podkladom pre nasledujúce odstavce je hlavne článok [Mu85].

Stať 2.2 obsahuje návrh modifikácie konštrukcie intervalových odhadov (algoritmus 1) z článku [WaGa98].

2.1 Jadrové odhady v prípade *i.i.d.*

Predpokladajme závislosť

$$Y_i = f(x_i) + e_i, \tag{2.1}$$

$e_i \sim i.i.d.$, $\mathsf{E}e_i = 0$, $\mathsf{var}\ e_i = \sigma^2$, $i = 1, \ldots, n$ a funkcia f nech je v nejakej "rozumnej" triede funkcií. Kritériom posudzovania kvality odhadu $\hat{f}$ funkcie f bude pre nás integrovaná L_2 vzdialenosť odhadu od skutočnej funkcie, teda

$$\mathsf{IMSE}(\hat{f}) = \mathsf{MISE}(\hat{f}) = \mathsf{E}\int [\hat{f}(x) - f(x)]^2 dx = \mathsf{IV} + \mathsf{ISB}. \tag{2.2}$$

IMSE=*integrated mean square error* a MISE=*mean integrated square error* sú vždy rovné (Fubiniova veta).

$$\mathsf{IV} = \int \mathsf{var}\ (\hat{f}(x))dx \ (\textit{integrated variance}), \tag{2.3}$$

$$\mathsf{ISB} = \int \mathsf{bias}^2(\hat{f}(x))dx \ (\textit{integrated squared bias}). \tag{2.4}$$

Funkciu f budeme odhadovať v tvare $f_n(x) = \sum_{i=1}^n w(i, x, \boldsymbol{x}, n)Y_i$, kde $\boldsymbol{x} = (x_1, \ldots, x_n)^T$. Funkcia w v nejakom zmysle "váži" hodnoty Y_i vzhľadom k tomu, v akom bode x funkciu f odhadujeme. Jej tvar určujeme samozrejme tak, aby sme dostali čo najrozumnejší odhad (globálne pre všetky x alebo lokálne). Väčšinou funkciu w hľadáme v nejakej rodine, vzhľadom k ďalšiemu parametru $b = b(n) = b_n$, poprípade $b_n(x)$. Tento parameter nazývame *bandwidth* (šírka pásma) alebo tiež vyhladzovací parameter (*smoothing parameter*).

Jadrovým odhadom nazývame taký odhad, keď $w = w(K)$, kde K je jadrová funkcia. Obecne možno za jadrovú funkciu považovať každú funkciu, ktorá je hustotou nejakého rozdelenia. Jadrový odhad je teda "lokálne vážený priemer".

Jadrové odhady regresnej funkcie sa dajú odvodiť z jadrových odhadov hustôt. Táto problematika je opísaná v [Sc92]. Nájdeme tu aj spomínané odvodenie

jadrového odhadu regresnej funkcie (Nadaraya-Watson) v kap. 8.1, podobne ako v [AnVo92]:

$$f_n^{NW}(x) = \sum_{i=1}^{n} \frac{K(\frac{x-x_i}{b})}{\sum_{j=1}^{n} K(\frac{x-x_j}{b})} Y_i. \tag{2.5}$$

Priestly a Chao navrhli pre $0 = x_1 < x_2 < \ldots < x_n = 1$, $f \in \mathcal{C}([0,1])$ odhad

$$f_n^{PC}(x) = \sum_{i=1}^{n} \frac{1}{b}(x_i - x_{i-1}) K\left(\frac{x - x_i}{b}\right) Y_i, \qquad x_0 = 0. \tag{2.6}$$

V našich úvahách budeme vychádzať z odhadu, ktorý pre $0 = x_1 < x_2 < \ldots < x_n = 1$ a $f \in \mathcal{C}([0,1])$ navrhli Gasser a Müller:

$$f_n^{GM}(x) = \sum_{i=1}^{n} \frac{1}{b} \int_{(x_i+x_{i-1})/2}^{(x_{i+1}+x_i)/2} K\left(\frac{x-t}{b}\right) dt \cdot Y_i, \qquad x_0 := 0, x_{n+1} := 1. \tag{2.7}$$

Odhady f_n^{PC} a f_n^{GM} sa dajú analogicky definovať na ľubovoľnom kompaktnom intervale $[A, B]$. Rôzne typy jadrových odhadov regresných funkcií popisuje napríklad článok [JoDaPa94].

Gasserov-Müllerov odhad (budeme značiť $f_{n,0}(x)$ a hovoriť GM odhad) má pre nás výhodné asymptotické aj teoretické vlastnosti, ako si ukážeme neskôr.

Pre $f \in \mathcal{C}^k([0,1])$ sa budeme zaoberať aj odhadom nulových bodov ν-tej derivácie ($\nu \leqq k-2$) regresnej funkcie f, jej odhad označíme $f_{n,\nu}$. Počíta sa "ν-tym zderivovaním" odhadu $f_{n,0}$. Namiesto jadrovej funkcie K bude v odhade vystupovať jadrová funkcia K_ν rádu ν (a k).

Definícia 2.1 (Jadrová funkcia rádu (ν, k)). Nech $\nu \geqq 0$, $k \geqq \nu + 2$ sú prirodzené čísla. Potom *jadrovou funkciou* rádu (ν, k) nazveme takú K_ν lipschitzovskú na [-1,1], pre ktorú

$$\int t^j K_\nu(t) dt = \begin{cases} 0 & 0 \leqq j < k \quad j \neq \nu \\ (-1)^\nu \nu! & j = \nu \\ (-1)^k k! B_k \neq 0 & j = k \end{cases} \tag{2.8}$$

Gasserov-Müllerov jadrový odhad ν-tej derivácie regresnej funkcie potom vyjde:

$$f_{n,\nu}(x) = \frac{1}{b^{\nu+1}} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_\nu\left(\frac{x-t}{b}\right) dt Y_i, \tag{2.9}$$

ak označíme $s_i := \frac{x_{i+1}+x_i}{2}$.

Predpoklady 2.2. V celom odstavci budeme uvažovať:

- K_ν s nosičom [-1,1];
- $b = b_n \xrightarrow{n\to\infty} 0$, $nb_n \xrightarrow{n\to\infty} \infty$;
- K_ν lipschitzovskú a nie identicky 0-vú na $\mathbb{R}$;
- $|x_i - x_{i-1} - n^{-1}| = o(n^{-1})$, $2 \leqq\ i \leqq n$;
- $f \in \mathcal{C}^k([0,1])$.

Ďalej označme

$$B_k := \frac{(-1)^k}{k!} \int_{-1}^{1} t^k K_\nu(t)dt; \tag{2.10}$$

$$V := \int_{-1}^{1} K_\nu^2(t)dt. \tag{2.11}$$

Lemma 2.3 (CLV pre GM odhad). *Za predpokladov 2.2 platí :*

$$\frac{f_{n,\nu}(x) - \mathsf{E} f_{n,\nu}(x)}{\sqrt{\mathsf{var}\ f_{n,\nu}(x)}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathsf{N}(0,1), \qquad \forall x \in [0,1]. \tag{2.12}$$

Dôkaz. Stačí použiť Lindebergovu CLV ([La03], str. 95, veta 17.4), Cramérovu-Sluckého vetu ([An02], str.337, veta B.10), princíp spojitosti pre konvergenciu v distribúcii ([La03], str. 73, veta 14.5) a vhodné integrálne aproximácie. □

Lemma 2.4 (Asymptotika vychýlenia jadrového odhadu). *Za predpokladov 2.2 pre $x \in [\delta, 1-\delta]$, $0 < \delta < \frac{1}{2}$ platí:*

$$\mathsf{E} f_{n,\nu}(x) - f^{(\nu)}(x) = b^{k-\nu}\{B_k f^{(k)}(x) + o(1)\} + \mathcal{O}\left(\frac{1}{nb^\nu}\right). \tag{2.13}$$

Dôkaz.

$$\mathsf{E} f_{n,\nu}(x) = \ldots = b^{-\nu} \int K_\nu(t) f(x - tb)dt + \mathcal{O}\left(\frac{1}{nb^\nu}\right) \tag{2.14}$$

$$\mathsf{var}\ f_{n,\nu}(x) = \ldots = \frac{\sigma^2}{nb^{2\nu+1}}(V + o(1)). \tag{2.15}$$

Taylorov rozvoj f v bode x do rádu k, vo vzorci (2.14). □

Toto vyjadrenie nám za ďalších predpokladov dáva silné tvrdenie, ktoré bude v celom odstavci esenciálne, a síce, že jadrový odhad konverguje k skutočnej funkcii f "rovnomerne skoro isto", teda je "*globálne konzistentný*". Formálne:

Veta 2.5 ("*Globálna konzistencia*" jadrového odhadu). *Nech platia predpoklady 2.2. Nech* $\mathsf{E}|e_1|^r < \infty$ *pre nejaké* $r > 2$ *a nech platí*

$$\liminf_{n\to\infty} nb^k > 0,\ \liminf_{n\to\infty} n^{1-2/r} b(\log n)^{-1} > 0.\ \textit{Potom}$$

$$\sup_{x\in[\delta,1-\delta]} |f_{n,\nu}(x) - f^{(\nu)}(x)| = \mathcal{O}\left\{ b^{k-\nu} + \left(\frac{\log n}{nb^{2\nu+1}}\right)^{1/2} \right\} \quad a.s., \qquad 0 < \delta < \frac{1}{2}. \tag{2.16}$$

Dôkaz. Vyplýva z (2.13) a lemmy 2 v [Mu84a]. □

Poznámka 2.6. Pri jadrových odhadoch s pevnou hodnotou (vzhľadom k x) vyhladzovacieho parametra b dochádza k tzv. "*boundary effectu*", t.j. veľkého vychýlenia pri krajoch intervalu $[0, 1]$. Existuje metóda, ktorá zaručí "globálnu konzistenciu" na celom intervale $[0, 1]$ (napr. tzv. "*smooth modified kernels*"), tá však na jeho krajoch náš odhad modifikuje. Podmienka asymptotickej ekvidistancie nie je takáto kľúčová, iba zjednodušuje výklad.

2.1.1 Odhady nulových bodov

Veta 2.5 nám dáva globálny vzťah medzi odhadom $f_{n,\nu}$ a skutočnou funkciou $f^{(\nu)}$. Možno teda očakávať, že pri splnení predpokladov bude možné na základe tejto vety odhadnúť určité vlastnosti skutočnej funkcie $f^{(\nu)}$ vlastnosťami funkcie $f_{n,\nu}$, a hlavne, štatisticky tieto odhady popísať.

Definícia 2.7 (Odhad nulového bodu). Predpokladajme, že $f^{(\nu)}$ má na intervale $(0, 1)$ práve jeden nulový bod ζ_ν. Potom definujeme odhad polohy nulového bodu

$$\zeta_{n,\nu} := \inf_{x\in[0,1]} \{f_{n,\nu}(x) = 0\}, \text{ dodefinujúc } \inf_{x\in[0,1]} \emptyset := 0; \tag{2.17}$$

V práci budeme porovnávať aj odhad polohy nulového bodu derivácie funkcie s odhadom extrému funkcie. Preto definujme ďalej:

Definícia 2.8 (Odhad extrému). Predpokladajme, že funkcia $f^{(\nu)}$ má na intervale $(0, 1)$ práve jedno maximum θ_ν. Potom definujeme odhad polohy maxima ako:

$$\theta_{n,\nu} = \inf_{x\in[0,1]} \{f_{n,\nu}(x) = \max_{t\in[0,1]} f_{n,\nu}(t)\}, \tag{2.18}$$

definícia odhadu polohy minima je analogická.

Nájdeme aj zaujímavý vzťah medzi odhadom polohy extrému $\theta_{n,\nu}$ a odhadom $f_{n,\nu}(\theta_{n,\nu})$ veľkosti extrému $f^{(\nu)}(\theta_\nu)$.

Korektné je ukázať, že takto definované "odhady" odhadmi skutočne sú.

Tvrdenie 2.9 (Merateľnosť jadrových odhadov). *Funkcie*

- $\zeta_{n,\nu}$ *pre* ζ_ν*;*
- $\theta_{n,\nu}$ *pre* θ_ν*;*
- $f_{n,\nu}(\theta_{n,\nu})$ *pre* $f^{(\nu)}(\theta_\nu)$

sú borelovsky merateľné, teda sú odhadmi.

Dôkaz. Podobne ako v [Ed80]. □

Na základe vety 2.5 budeme chcieť ukázať konzistenciu našich odhadov. Pre všetky nasledujúce vety budeme predpokladať, že

$$\sup_{x\in[\delta,1-\delta]} |f_{n,\nu}(x) - f^{(\nu)}(x)| = \mathcal{O}(\beta_n) \quad a.s., \tag{2.19}$$

kde $\beta_n \xrightarrow{n\to\infty} 0$. Tento predpoklad je splniteľný, prihliadnuc k vete 2.5 môžeme voliť

$$\beta_n := \mathcal{O}\left(\frac{(\log n)^{1/2}}{n^{\frac{k-\nu}{2k+1}}}\right), \text{ ak zvolíme } b := \mathcal{O}\left(n^{-\frac{1}{2k+1}}\right). \tag{2.20}$$

Aby však ostali aj ostatné predpoklady vety 2.5 splnené, musí platiť

$$\begin{aligned} \liminf_{n\to\infty} n^{1-2/r} b(\log n)^{-1} &> 0 \Leftrightarrow \\ \Leftrightarrow 1 - \frac{2}{r} - \frac{1}{2k+1} &> 0 \Leftrightarrow \\ \Leftrightarrow r &> 2 + \frac{1}{k}, \end{aligned}$$

takže musí byť $\mathsf{E}|e_1|^r < \infty$ pre nejaké $r > 2 + \frac{1}{k}$.

Takto môžeme pristúpiť k nasledujúcim tvrdeniam.

Lemma 2.10 (Konzistencia odhadu nulového bodu). *Nech platí predpoklad* (2.19) *a predpoklady 2.2. Nech existujú* a_0*,* b_0*,* c *a* τ *také, že* $0 < a_0 < \zeta_\nu < b_0 < 1$*,* $c > 0$*,* $\tau \geqq 1$ *a* $f^{(\nu)}$ *je rýdzomonotónna na* $[a_0, b_0]$ *a spĺňuje* $|f^{(\nu)}(x)| > c|x - \zeta_\nu|^\tau$ *pre* $x \in [a_0, b_0]$*,* $x \neq \zeta_\nu$*. Potom*

$$|\zeta_{n,\nu} - \zeta_\nu| = \mathcal{O}(\beta_n^{1/\tau}) \quad a.s. \tag{2.21}$$

Dôkaz. Pretože $f^{(\nu)}$ je rýdzomonotónna a spojitá a ζ_ν je jej jediný nulový bod, existuje $\delta > 0$ taká, že $|f^\nu(x)| > \delta$ pre $x \notin [a_0, b_0]$, pričom $sgn(f^{(\nu)}(x))$ je rôzne pre $x < a_0$ a $x > b_0$. Pre n veľké je $\beta_n < \frac{\delta}{2}$ a vzhľadom k (2.19) aj $|f_{n,\nu}(x)| > \frac{\delta}{2}$ *a.s.* pre $x \notin [a_0, b_0]$ a $sgn(f_{n,\nu}(x))$ je *a.s.* rôzne pre $x < a_0$ a $x > b_0$. Z toho vyplýva, že $\zeta_{n,\nu} \in [a_0, b_0]$ *a.s.* a odtiaľto, že

$$|\zeta_{n,\nu} - \zeta_\nu|^\tau < \frac{1}{c}|f^{(\nu)}(\zeta_{n,\nu})| = \frac{1}{c}|f^{(\nu)}(\zeta_{n,\nu}) - f_{n,\nu}(\zeta_{n,\nu})| = \mathcal{O}(\beta_n) \; a.s.$$

□

Podobne pre odhady polohy a veľkosti extrému.

Lemma 2.11 (Konzistencia odhadov polohy a veľkosti extrému). *Nech platí predpoklad* (2.19) *a predpoklady 2.2. Nech existujú* a_0, b_0, c *a* ρ *také, že* $0 < a_0 < \zeta_\nu < b_0 < 1$, $c > 0$, $\rho \geqq 1$ *a* $f^{(\nu)}$ *je rýdzorastúca (rýdzoklesajúca) na* $[a_0, \theta_\nu]$ *a rýdzoklesajúca (rýdzorastúca) na* $[\theta_\nu, b_0]$ *a splňujúca* $|f^{(\nu)}(x) - f^{(\nu)}(\theta_\nu)| > c|x - \theta_\nu|^\rho$ *pre* $x \in [a_0, b_0]$, $x \neq \theta_\nu$. *Potom*

$$|\theta_{n,\nu} - \theta_\nu| = \mathcal{O}(\beta_n^{1/\rho}) \quad a.s. \quad a \quad |f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu)| = \mathcal{O}(\beta_n) \quad a.s. \tag{2.22}$$

Dôkaz. Bez újmy na obecnosti uvažujme maximá.
Pretože $f^{(\nu)}$ má jediné maximum v bode θ_ν, existuje $\delta > 0$ tak, že $f^{(\nu)}(\theta_\nu) > f^{(\nu)}(x) + \delta$ pre $x \notin [a_0, b_0]$. Pre dostatočne veľké n je $\beta_n < \frac{\delta}{2}$, preto vzhľadom k (2.19) platí $f_{n,\nu}(\theta_\nu) > f^{(\nu)}(\theta_\nu) - \frac{\delta}{2} > f_{n,\nu}(x)$ *a.s.* pre $x \notin [a_0, b_0]$, odkiaľ vyplýva, že aj $\theta_{n,\nu} \in [a_0, b_0]$ *a.s.* Potom

$$\begin{aligned}
|\theta_{n,\nu} - \theta_\nu|^\rho &< \frac{1}{c}\left(f^{(\nu)}(\theta_\nu) - f^{\nu}(\theta_{n,\nu})\right) \\
&\leqq \frac{1}{c}\left(f^{(\nu)}(\theta_\nu) - f^{(\nu)}(\theta_{n,\nu}) + f_{n,\nu}(\theta_{n,\nu}) - f_{n,\nu}(\theta_\nu)\right) \\
&\leqq \frac{1}{c}\left(|f^{(\nu)}(\theta_\nu) - f_{n,\nu}(\theta_\nu)| + |f^{(\nu)}(\theta_{n,\nu}) - f_{n,\nu}(\theta_{n,\nu})|\right) = \mathcal{O}(\beta_n) \; a.s.
\end{aligned}$$

Pre veľkosť maxím

$$\begin{aligned}
|f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu)| &= |f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu) + f^{(\nu)}(\theta_{n,\nu}) - f^{(\nu)}(\theta_{n,\nu})| \\
&\leqq \left(f^{(\nu)}(\theta_\nu) - f^{(\nu)}(\theta_{n,\nu})\right) + \left|f^{(\nu)}(\theta_{n,\nu}) - f_{n,\nu}(\theta_{n,\nu})\right| \\
&= \mathcal{O}(\beta_n) \; a.s.
\end{aligned}$$

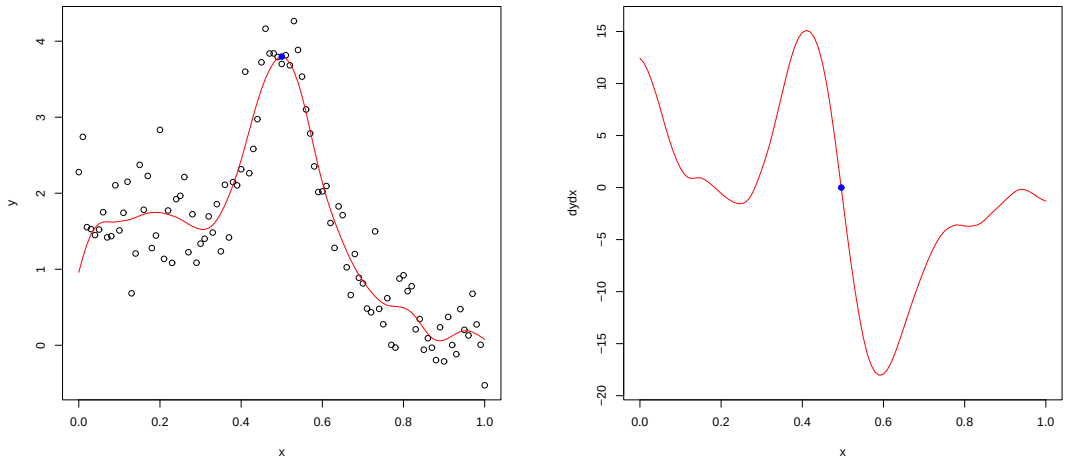
podľa časti dôkazu pre polohu maxím. □

Príklad 2.12 (Odhad polohy extrému a nulového bodu prvej derivácie). Skúmame dáta $(x_1, Y_1), \ldots (x_{101}, Y_{101})$ a predpokladáme model:

$$Y_i = f(x_i) + e_i; \quad e_i \sim \mathsf{N}(0; \sigma^2); \quad i = 1, \ldots, 101.$$

V skutočnosti $f(x) = 2 - 2x + 3\exp\left\{-\left(\frac{x-0.5}{0.1}\right)^2\right\}$, $\theta_0 = 0.4967 = \zeta_1$, $\sigma^2 = 0.2$.

Na obr. 6 sú znázornené odhady regresnej funkcie a prvej derivácie a sú vyznačené odhady $\theta_{101,0} = 0.500$ a $\zeta_{101,1} = 0.496$.



Obr. 6: Jadrový odhad funkcie a prvej derivácie

Pre odhad funkcie je použitý vyhladzovací parameter $b = 0.086$, ktorý bol určený metódou *cross-validation* (vzorec (2.55) a príklad 2.20) a jadrová funkcia rádu (0,2) (veta 2.19, $\mu = 3$). Pre odhad derivácie je zvolené $b = 0.144$ a jadrová funkcia rádu (1,3) (veta 2.19, $\mu = 2$). Ešte pripomeňme, že za odhad funkcie, resp. derivácie na krajoch $[0, b)$ a $(1 - b, 1]$ intervalu $[0, 1]$ neručíme. Ďalej priznávame, že sme odhad nulového bodu derivácie pre odhad maxima nepočítali na základe definície, ale zobrali sme ten nulový bod odhadu derivácie, ktorý je najbližšie polohe maxima odhadu pôvodnej funkcie.

Poznámka 2.13. Príklad 2.12 je prevzatý z článku [Mu85], s opravou o skutočnú hodnotu polohy extrému θ_0, o ktorej sa autor článku domnieval, že je rovná $\theta_0 = 0.5$.

2.1.2 Limitné rozdelenia

V tejto stati si ukážeme asymptotickú normalitu našich odhadov, odkiaľ je možné odvodiť približné intervalové odhady.

Pamätajme na značenia (2.10) a (2.11) a označme ešte

$$V' := \int K_\nu'^2(x)dx \tag{2.23}$$

Dôkazy nasledujúcich troch tvrdení možno nájsť aj v článku [Mu85]. Tu sú uvedené v mierne poupravenej a poopravenej forme, avšak najmä pre ich názornosť.

Veta 2.14 (CLV pre odhad nulového bodu). *Nech pre nejaké $r > 2$ je $\mathsf{E}|e_1|^r < \infty$ a nech*

$$\liminf_{n\to\infty} n^{1-2/r} b (\log n)^{-1} > 0.$$

Predpokladajme ďalej, že

$$f \in \mathcal{C}^{k+1}([0,1]), \quad \liminf_{n\to\infty} nb^{k+1} > 0 \quad a \quad \frac{nb^{2\nu+3}}{\log n} \xrightarrow{n\to\infty} \infty.$$

Nech K_ν je diferencovateľná, K'_ν je lipschitzovská na $\mathbb{R}$, nech $f^{(\nu+1)}(\zeta_\nu) \neq 0$ a nech sú splnené predpoklady lemmy 2.10. Ak pre nejaké $d \geqq 0$ pri $n \to \infty$ platí $nb^{2k+1} \to d^2$, potom

$$(nb^{2\nu+1})^{\frac{1}{2}} (\zeta_{n,\nu} - \zeta_\nu) \xrightarrow[n\to\infty]{\mathcal{D}} \mathsf{N}\left(-\frac{df^{(k)}(\zeta_\nu) B_k}{f^{(\nu+1)}(\zeta_\nu)}, \frac{\sigma^2 V}{(f^{(\nu+1)}(\zeta_\nu))^2} \right). \tag{2.24}$$

Dôkaz. Z požiadaviek na jadrovú funkciu K_ν vyplýva

$$\int K'_\nu(x) x^j dx = \begin{cases} 0 & j = 0, \ldots, k; j \neq \nu + 1 \\ (-1)^{\nu+1}(\nu+1)! & j = \nu + 1 \end{cases}$$

Keďže podľa predpokladov vety je K'_ν lipschitzovská na $\mathbb{R}$, sú splnené predpoklady vety 2.5 pre (ν, k) nahradené $(\nu + 1, k + 1)$, podľa ktorej

$$\sup_{x\in[\delta,1-\delta]} |f'_{n,\nu}(x) - f^{(\nu+1)}(x)| \to 0 \ a.s. \tag{2.25}$$

Z Lagrangeovej vety vyplýva, že existuje taký bod $\zeta^*_{n,\nu}$ ležiaci medzi bodmi ζ_ν a $\zeta_{n,\nu}$, pre ktorý

$$0 = f_{n,\nu}(\zeta_{n,\nu}) = f_{n,\nu}(\zeta_\nu) + (\zeta_{n,\nu} - \zeta_\nu) f'_{n,\nu}(\zeta^*_{n,\nu}).$$

Odtiaľto pomocou (2.25) a lemmy 2.10, podľa ktorej $\zeta_{n,\nu} \to \zeta_\nu$, a tak aj $\zeta^*_{n,\nu} \to \zeta_\nu$ a zo spojitosti funkcie $f^{(\nu+1)}$ vyplýva, že

$$\begin{aligned} & |f'_{n,\nu}(\zeta^*_{n,\nu}) - f^{(\nu+1)}(\zeta_\nu)| \leqq \\ \leqq\ & |f'_{n,\nu}(\zeta^*_{n,\nu}) - f^{(\nu+1)}(\zeta^*_{n,\nu})| + |f^{(\nu+1)}(\zeta^*_{n,\nu}) - f^{(\nu+1)}(\zeta_\nu)| \to 0 \ a.s. \end{aligned}$$

Pretože $f^{(\nu+1)}(\zeta_\nu) \neq 0$, aj $f'_{n,\nu}(\zeta^*_{n,\nu}) \neq 0$ *a.s.* pre dostatočne veľké n. Takto dostávame

$$\zeta_{n,\nu} - \zeta_\nu = \frac{f^{(\nu)}(\zeta_\nu) - f_{n,\nu}(\zeta_\nu)}{f'_{n,\nu}(\zeta^*_{n,\nu})}.$$

Výsledok potom hneď plynie z predchádzajúcich úprav, lemmy 2.3 o asymptotickej normalite, vzorcov (2.13) o vychýlení a (2.15) o rozptyle jadrového odhadu, z princípu spojitosti pre konvergenciu v distribúcii a Cramér-Sluckého vety. □

Veta 2.15 (CLV pre odhad polohy extrému). *Nech pre nejaké $r > 2$ je $\mathsf{E}|e_1|^r < \infty$ a nech*

$$\liminf_{n\to\infty} n^{1-2/r} b (\log n)^{-1} > 0.$$

Predpokladajme ďalej, že

$$f \in \mathcal{C}^{k+2}([0,1]), \quad \liminf_{n\to\infty} nb^{k+2} > 0 \quad a \quad \frac{nb^{2\nu+5}}{\log n} \xrightarrow{n\to\infty} \infty.$$

Nech K_ν je dvakrát diferencovateľná, K_ν'' je lipschitzovská na $\mathbb{R}$, nech $f^{(\nu+2)}(\theta_\nu) \neq 0$ a nech sú splnené predpoklady lemmy 2.11. Ak pre nejaké $d' \geqq 0$ pri $n \to \infty$ platí $nb^{2k+3} \to d'^2$, potom

$$(nb^{2\nu+3})^{\frac{1}{2}} (\theta_{n,\nu} - \theta_\nu) \xrightarrow[n\to\infty]{\mathcal{D}} \mathsf{N}\left(-\frac{d' f^{(k+1)}(\theta_\nu) B_k}{f^{(\nu+2)}(\theta_\nu)}, \frac{\sigma^2 V'}{(f^{(\nu+2)}(\theta_\nu))^2} \right). \tag{2.26}$$

Dôkaz. Podobne ako v predchádzajúcej vete je možné ukázať, že podmienky tejto vety zaručujú konvergenciu

$$\sup_{x\in[\delta,1-\delta]} |f''_{n,\nu}(x) - f^{(\nu+2)}(x)| \to 0 \ a.s. \tag{2.27}$$

podľa vety 2.5. Potom pomocou lemmy 2.11 platí

$$|f''_{n,\nu}(\theta_{n,\nu}) - f^{(\nu+2)}(\theta_\nu)| \to 0 \ a.s., \tag{2.28}$$

odkiaľ vyplýva, že pre dostatočne veľké n je $f''_{n,\nu}(\theta_{n,\nu}) \neq 0$ $a.s.$ Potom

$$\theta_{n,\nu} - \theta_\nu = \frac{f^{(\nu+1)}(\theta_\nu) - f'_{n,\nu}(\theta_\nu)}{f^{(\nu+2)}(\theta_\nu)} + R_n \tag{2.29}$$

podľa Lagrangeovej vety o strednej hodnote a pre nejaké $\theta^*_{n,\nu}$ medzi θ_ν a $\theta_{n,\nu}$ platí rovnosť

$$R_n = \frac{\left(f^{(\nu+1)}(\theta_\nu) - f'_{n,\nu}(\theta_\nu)\right)\left(f^{(\nu+2)}(\theta_\nu) - f''_{n,\nu}(\theta^*_{n,\nu})\right)}{f^{(\nu+2)}(\theta_\nu) \cdot f''_{n,\nu}(\theta^*_{n,\nu})}. \tag{2.30}$$

Vzhľadom k predpokladom kladeným na jadrovú funkciu je

$$\frac{(-1)^{k+1}}{(k+1)!} \int x^{k+1} K'_\nu(x) dx = B_k. \tag{2.31}$$

A pretože K'_ν spĺňuje definíciu 2.1 pre (ν, k) zamenené postupne za $(\nu+1, k+1)$, dostávame

$$\mathsf{var}\ f'_{n,\nu}(x) = \frac{\sigma^2}{nb^{2\nu+3}} \left(V' + o(1)\right) \tag{2.32}$$

a tiež

$$\mathsf{E} f'_{n,\nu}(x) - f^{(\nu+1)}(x) = b^{k-\nu} \left(B_k f^{(k+1)}(x) + o(1)\right) + \mathcal{O}\left(\frac{1}{nb^{\nu+1}}\right). \tag{2.33}$$

Užitím lemmy 2.3 na funkciu $f'_{n,\nu}$ a pretože (2.28) zaručuje

$$(nb^{2\nu+3})^{\frac{1}{2}} R_n \to 0 \ a.s., \tag{2.34}$$

dostávame žiadaný výsledok podobne ako v závere dôkazu predchádzajúcej vety. □

Poznámka 2.16. Je dobré si povšimnúť porovnaním viet 2.14 a 2.15, že je asymptoticky ekvivalentné, či odhadujeme polohu nulového bodu funkcie $f^{(\nu+1)}$ alebo polohu extrému funkcie $f^{(\nu)}$, čo iba podporuje konzistentnosť celej teórie.

Veta 2.17 (CLV pre združený odhad polohy a veľkosti extrému). *Nech pre nejaké $r > 2$ je $\mathsf{E}|e_1|^r < \infty$ a nech*

$$\liminf_{n\to\infty} n^{1-2/r} b (\log n)^{-1} > 0.$$

Predpokladajme ďalej, že

$$f \in \mathcal{C}^{k+2}([0,1]), \quad \liminf_{n\to\infty} nb^{k+2} > 0 \quad a \quad \frac{nb^{2\nu+5}}{\log n} \xrightarrow{n\to\infty} \infty.$$

Nech K_ν je dvakrát diferencovateľná, K''_ν je lipschitzovská na $\mathbb{R}$, nech $f^{(\nu+2)}(\theta_\nu) \neq 0$ a nech sú splnené predpoklady lemmy 2.11. Ak pre nejaké $d \geqq 0$ pri $n \to \infty$ platí $nb^{2k+1} \to d^2$ (tvrdenie platí pre $k > \nu + 2$), potom

$$\begin{aligned}&[(nb^{2\nu+3})^{\frac{1}{2}}(\theta_{n,\nu} - \theta_\nu), \quad (nb^{2\nu+1})^{\frac{1}{2}}\{f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu)\}] \\ &\xrightarrow[n\to\infty]{\mathcal{D}} \mathsf{N}_2 \left\{ \begin{pmatrix} 0 \\ df^{(k)}(\theta_\nu)B_k \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2 V'}{(f^{(\nu+2)}(\theta_\nu))^2} & 0 \\ 0 & \sigma^2 V \end{pmatrix} \right\}. \end{aligned} \tag{2.35}$$

Dôkaz. Ukážeme pomocou vety 2.15, lemmy 2.3 a Cramér-Woldovho princípu pre konvergenciu náhodného vektoru v distribúcii, viď [La03], str. 86, veta 15.24. Potrebujeme ukázať, že $\forall(\lambda, \eta) \in \mathbb{R}^2$ platí:

$$\lambda(nb^{2\nu+3})^{1/2}(\theta_{n,\nu} - \theta_\nu) + \eta(nb^{2\nu+1})^{1/2}\left(f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu)\right) \tag{2.36}$$

$$\xrightarrow[n\to\infty]{\mathcal{D}} \mathsf{N}\left(\eta d f^{(k)}(\theta_\nu)B_k, \lambda^2 \frac{\sigma^2 V'}{\left(f^{(\nu+2)}(\theta_\nu)\right)^2} + \eta^2\sigma^2 V\right). \tag{2.37}$$

Vzhľadom k (2.13) pri $n \to \infty$

$$(nb^{2\nu+1})^{1/2}\left(\mathsf{E}f_{n,\nu}(\theta_\nu) - f^{(\nu)}(\theta_\nu)\right) \to df^{(k)}(\theta_\nu)B_k. \tag{2.38}$$

Podľa vety 2.15

$$\mathcal{L}\left\{(nb^{2\nu+3})^{1/2}(\theta_{n,\nu} - \theta_\nu)\right\} \to \mathsf{N}\left(0, \frac{\sigma^2 V'}{(f^{(\nu+2)}(\theta_\nu))^2}\right), \tag{2.39}$$

pretože $(nb^{2k+3} \to 0) \Leftarrow (nb^{2k+1} \to d^2)$ pre $0 \leqq d \in \mathbb{R}$ pri $n \to \infty$.
Z predpokladov kladených na b a z (2.39) vyplýva, že

$$\frac{1}{b}(\theta_{n,\nu} - \theta_\nu) \xrightarrow[n\to\infty]{P} 0. \tag{2.40}$$

Pre vhodné $\theta^*_{n,\nu}$ medzi $\theta_{n,\nu}$ a θ_ν a pre vhodné $\theta^{**}_{n,\nu}$ medzi $\theta_{n,\nu}$ a $\theta^*_{n,\nu}$ je

$$f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu) = f_{n,\nu}(\theta_\nu) + f'_{n,\nu}(\theta^*_{n,\nu})(\theta_{n,\nu} - \theta_\nu) - f^{(\nu)}(\theta_\nu)$$

a

$$f'_{n,\nu}(\theta^*_{n,\nu}) = f'_{n,\nu}(\theta^*_{n,\nu}) - f'_{n,\nu}(\theta_{n,\nu}) = f''_{n,\nu}(\theta^{**}_{n,\nu})(\theta^*_{n,\nu} - \theta_{n,\nu}),$$

čo vyplýva z toho, že $|\theta^*_{n,\nu} - \theta_{n,\nu}| < |\theta_\nu - \theta_{n,\nu}|$, 2.40 a vzťahu

$$f''_{n,\nu}(\theta^{**}_{n,\nu}) \xrightarrow[n\to\infty]{P} f^{(\nu+2)}(\theta_\nu) \neq 0,$$

podľa (2.28). Odtiaľto plynie, že

$$\frac{1}{b} f'_{n,\nu}(\theta^*_{n,\nu}) \xrightarrow[n\to\infty]{P} 0.$$

Ďalej pomocou (2.39) dostávame

$$(nb^{2\nu+1})^{1/2} f'_{n,\nu}(\theta^*_{n,\nu})(\theta_{n,\nu} - \theta_\nu) \xrightarrow[n\to\infty]{P} 0.$$

Spojením s výsledkom (2.38) máme

$$(nb^{2\nu+1})^{1/2} \left(f_{n,\nu}(\theta_{n,\nu}) - f^{(\nu)}(\theta_\nu)\right) =$$

$$= (nb^{2\nu+1})^{1/2} \left(f_{n,\nu}(\theta_\nu) - \mathsf{E} f_{n,\nu}(\theta_\nu)\right) + df^{(k)}(\theta_\nu) B_k + o_P(1). \tag{2.41}$$

Platí $(nb^{2\nu+3})^{1/2} \left(f^{(\nu+1)}(\theta_\nu) - \mathsf{E} f'_{n,\nu}(\theta_\nu)\right) \to 0$ podľa (2.33) a pretože $nb^{2k+3} \to 0$. Potom, vzhľadom k (2.34), dostávame

$$(nb^{2\nu+3})^{1/2}(\theta_{n,\nu} - \theta_\nu) = -(nb^{2\nu+3})^{1/2} \frac{\left(f'_{n,\nu}(\theta_\nu) - \mathsf{E} f'_{n,\nu}(\theta_\nu)\right)}{f^{(\nu+2)}(\theta_\nu)} + o_P(1). \tag{2.42}$$

Uvážením výsledkov (2.41) a (2.42) po dosadení tvaru jadrového odhadu a použitím Lebesgueovej vety vidíme, že asymptotické rozdelenie (2.36) je rovnaké ako asymptotické rozdelenie veličiny:

$$(nb^{2\nu+1})^{1/2} \left[\frac{1}{b^{\nu+1}} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} \left\{ -\frac{\lambda}{f^{(\nu+2)}(\theta_\nu)} K'_\nu \left(\frac{\theta_\nu - t}{b} \right) + \eta K_\nu \left(\frac{\theta_\nu - t}{b} \right) \right\} dt \cdot e_i \right]$$

$$+ \quad \eta d f^{(k)}(\theta_\nu) B_k. \tag{2.43}$$

Za povšimnutie stojí rozšírenie zlomku $\frac{(nb^{2\nu+1})^{1/2}}{b^{\nu+1}}$ číslom b na $\frac{(nb^{2\nu+3})^{1/2}}{b^{\nu+2}}$ pre prvý sčítanec "vnútri" integrálu. Uplatnením CLV pre jadrové odhady (veta 2.3) na (2.43), substitúciou $u := \frac{\theta_\nu - t}{b}$ a využitím aditívnej vlastnosti hraníc integrálu dostávame asymptotické rozdelenie

$$\mathsf{N}\left(\eta d f^{(k)}(\theta_\nu) B_k, \sigma^2 \int_{-1}^{1} \left\{ \frac{\lambda^2 K_\nu'^2(u)}{\left(f^{(\nu+2)}(\theta_\nu)\right)^2} - \frac{2\lambda\eta K_\nu'(u) K_\nu(u)}{f^{(\nu+2)}(\theta_\nu)} + \eta^2 K_\nu^2(u) \right\} du \right).$$

Pretože

$$\int_{-1}^{1} K_\nu(u) K_\nu'(u) du = 0, \tag{2.44}$$

je tvrdenie vety dokázané. □

Poznámka 2.18. V článku [WaGa98] sa dajú nájsť zobecnenia (veta 3.1 a tvrdenie 3.1 na str. 979) tvrdení ostatných troch viet, ktoré hovoria o limitných rozdeleniach (normálnych) odhadov (odhadu) bodov so všeobecnejšími vlastnosťami (vo vzťahu k funkcii f).

Aby sme na základe týchto viet dosiahli dobré bodové a čo najlepšie približné intervalové odhady, musíme jednak konzistentne odhadnúť parameter σ^2 (tomu sa venuje kapitola 2.1.3) a hodnoty $f^{(\nu+1)}(\zeta_\nu)$, resp. $f^{(\nu+2)}(\theta_\nu)$ (odhadneme na základe vety 2.5 pomocou vzťahov (2.58) alebo (2.59)), jednak čo najlepšie zvoliť jadrovú funkciu K_ν (kapitola 2.1.4) a, najmä a hlavne (!), vyhladzovací parameter b (kapitola 2.1.5), ktorý bude spĺňať predpoklady príslušnej vety.

2.1.3 $\widehat{\sigma^2}$

Konzistentný odhad rozptylu dostaneme jednoducho ako

$$\widehat{\sigma^2} = \frac{1}{n}\mathsf{RSS} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}(x_i)\right)^2. \tag{2.45}$$

Tento odhad však príliš závisí na tom, ako odhadneme funkciu f, resp., akú jadrovú funkciu a aký vyhladzovací parameter b použijeme. Niektoré metódy voľby vyhladzovacieho parametra (kapitola 2.1.5) potrebujú mať spočítaný odhad rozptylu vopred. Preto sa používajú aj iné odhady, napr. metóda *three points residuals*:

$$\widehat{\sigma^2} = \frac{2}{3(n-2)}\sum_{i=2}^{n-1}\left(Y_i - \frac{Y_{i-1} + Y_{i+1}}{2}\right)^2. \tag{2.46}$$

2.1.4 Jadrové funkcie

V tomto odstavci sa budeme zaoberať voľbou funkcií K_ν. Voľba jadrovej funkcie nie je pre konečný odhad taká kľúčová ako bude voľba vyhladzovacieho parametra; každopádne, ak budeme chcieť používať vety 2.14, 2.15 alebo 2.17, budeme

musieť splniť ich predpoklady, a to aj o jadrovej funkcii K_ν. Je zrejmé, že máme na výber nekonečne mnoho funkcií, ktoré spĺňajú definíciu 2.1 jadrovej funkcie K_ν rádu (ν, k). Preto môžeme medzi týmito funkciami hľadať z hľadiska IMSE funkciu optimálnu pri nejakom pevnom b. Tento problém je aj tak ešte stále príliš obecný, preto sa budeme snažiť o minimalizáciu iba jednej zložky (IV, a tá je funkciou (rastúcou v) $\int K_\nu^2(x)dx$ vzhľadom k jadrovej funkcii, teda

$$\min_{K_\nu \in \mathcal{M}_{\nu,k}} \int K_\nu^2(x)dx, \tag{2.47}$$

kde $\mathcal{M}_{\nu,k}$ je priestor všetkých jadrových funkcií rádu (ν, k) z priestoru $L^2([-1,1])$, v tejto stati predpokladajme ν, k s rovnakou paritou. Riešenie tohoto problému však vedie k jadrám nespojitým v bodoch -1 a 1 a odhadovaným funkciám $f_{n,\nu}$ nediferencovateľným. Preto zavádzame ďalší parameter μ, ktorý popisuje hladkosť jadrovej funkcie $K_\nu \in \mathcal{C}^\mu([-1,1])$ a zároveň očakávanú hladkosť odhadnutej funkcie $f_{n,\nu} \in \mathcal{C}^\mu([0,1])$, ak pridáme predpoklad

$$K_\nu^{(j)}(-1) = K_\nu^{(j)}(1) = 0, \quad j = 0, \ldots, \mu - 1. \tag{2.48}$$

Optimalizačná úloha potom nadobudne tvar:

$$\min_{K_\nu \in \mathcal{M}_{\nu,k} \cap \mathcal{C}^\mu([-1,1])} \int {K_\nu^{(\mu)}}^2(x)dx \quad \text{za podmienky (2.48).} \tag{2.49}$$

Veta 2.19 (Smooth Optimum Kernels). *Optimalizačná úloha* (2.49) *má práve jedno riešenie, a tým je reštrikcia polynómu stupňa* $(k + 2\mu - 2)$ *na interval* $[-1, 1]$, *pritom koeficienty* γ_i $(0 \leqq i \leqq k + 2\mu - 2)$ *pri* x^i *sú dané vzťahom*

$$\gamma_i = \begin{cases} \frac{(-1)^{(i+\nu)/2}(k+\nu+2\mu)!(k+i)!(k-\nu)(k+2\mu-i)}{i!(i+\nu+1)2^{2(k+\mu)+1}\left(\frac{k-\nu}{2}\right)!\left(\frac{k+\nu+2\mu}{2}\right)!\left(\frac{k+2\mu-i}{2}\right)!\left(\frac{k+i}{2}\right)!} & \textit{pre } (k+i) \textit{ párne} \\ 0 & \textit{pre } (k+i) \textit{ nepárne} \end{cases}. \tag{2.50}$$

Dôkaz. viď [Mu84b], veta 2.4., str. 771. □

Niektoré optimálne jadrové funkcie sú v tab. 1:

(ν, k, μ)	K_ν	V	V'	B_k
$(0,2,1)$	$\frac{3}{4} - \frac{3}{4}x^2$	0.600	1.500	0.100
$(0,2,2)$	$\frac{15}{16} - \frac{15}{8}x^2 + \frac{15}{16}x^4$	0.714	2.143	0.071
$(0,4,1)$	$\frac{45}{32} - \frac{75}{16}x^2 + \frac{105}{32}x^4$	1.250	9.375	−0.002
$(0,4,2)$	$\frac{105}{64} - \frac{525}{64}x^2 + \frac{735}{64}x^4 - \frac{315}{64}x^6$	1.407	11.932	−0.001
$(1,3,1)$	$-\frac{15}{4}x + \frac{15}{4}x^3$	2.143	22.500	0.071
$(1,5,1)$	$-\frac{525}{32}x + \frac{735}{16}x^3 - \frac{945}{32}x^5$	11.932	275.625	−0.001
$(2,4,1)$	$-\frac{105}{16} + \frac{315}{8}x^2 - \frac{525}{16}x^4$	35.000	787.500	0.056
$(2,6,1)$	$-\frac{1575}{64} + \frac{19845}{64}x^2 - \frac{42525}{64}x^4 + \frac{24255}{64}x^6$	381.635	15946.880	−0.001

Tab. 1: Prehľad niektorých jadrových funkcií a ich parametre

2.1.5 Vyhladzovací parameter

Ako sme už spomínali, voľba vyhladzovacieho parametra je pre jadrové odhady absolútne podstatná. Predpokladajme, že hľadáme odhad $f_{n,\nu} = \widehat{f^{(\nu)}}$ funkcie $f^{(\nu)}$. Budeme sa snažiť o nájdenie optimálneho vyhladzovacieho parametra b (obmedzme sa na prípad "*global bandwidth*", teda $b = b_n$ pevné pre celý interval, na ktorom odhadujeme) minimalizáciou $\mathsf{IMSE} = \mathsf{IV} + \mathsf{ISB}$.

$$\widehat{f^{(\nu)}}(x) = \sum_{i=1}^{n} w_i Y_i, \tag{2.51}$$

$$\mathsf{var}\ \widehat{f^{(\nu)}}(x) = \sigma^2 \sum_{i=1}^{n} w_i^2, \tag{2.52}$$

$$\mathsf{bias}\left(\widehat{f^{(\nu)}}(x)\right) = \sum_{i=1}^{n} w_i f(x_i) - f^{(\nu)}(x), \tag{2.53}$$

odkiaľ spočítame pri známych f, σ^2 presne $\mathsf{IMSE}(b)$. K spočítaniu IMSE teda potrebujeme funkciu f a rozptyl chýb σ^2, ktoré však nemáme.

Voľba b pre odhad regresnej funkcie

Ak budeme hľadať odhad $f_{n,0} = \hat{f}$ priamo regresnej funkcie f, ako náhrada minimalizácie IMSE sa nám ponúka minimalizácia

$$\mathsf{RSS}(b) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{f}(x_i)\right)^2. \tag{2.54}$$

Táto metóda však príliš vyžaduje fit dát, čo vzhľadom k minimalizácii IMSE nemusí byť najdôležitejšie. Preto túto metódu modifikujeme:

$$b_{CV} = \arg\min_b \mathsf{CV}(b) = \arg\min_b \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{f_{-i}}(x_i)\right)^2, \tag{2.55}$$

kde $\widehat{f_{-i}}(x_i)$ je odhad funkcie f v bode x_i našou metódou pri pevne volenom b založený na $(x_1, Y_1), \ldots, (x_{i-1}, Y_{i-1}), (x_{i+1}, Y_{i+1}), \ldots, (x_n, Y_n)$. Tento postup sa nazýva *cross-validation* (krížové overovanie).

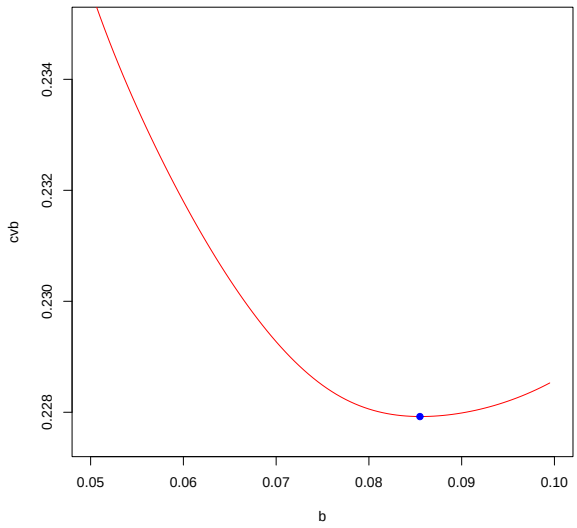
Ďalšou možnosťou voľby parametru b je tzv. *Riceovo kritérium*:

$$\mathsf{R}(b) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{f}(x_i)\right)^2 - \widehat{\sigma^2} + \frac{2\widehat{\sigma^2}}{nb} K(0), \tag{2.56}$$

kde $\widehat{\sigma^2}$ je konzistentný odhad σ^2, napr. metódou *three points residuals* - vzorec (2.46).

Minimalizáciou funkcie $\mathsf{R}(b)$, resp. $\mathsf{CV}(b)$ dostaneme konzistentný odhad b_R, resp. b_{CV} parametra b vzhľadom k IMSE, v zmysle (2.60). Výhodou *Riceovho kritéria* je, že nenapočítava nové odhady $\widehat{f_{-i}}$. Jeho nevýhodou oproti *cross-validation* je požiadavka existencie ôsmeho momentu rozdelenia chýb. V nasledujúcej stati spomenuté metódy zovšeobecíme a popíšeme ich z teoretického hľadiska dôkladnejšie.

Príklad 2.20 (Cross-validation). Skúmame dáta z príkladu 2.12. Metódou cross-validácie hľadáme odhad optimáleho vyhladzovacieho parametra pre odhad regresnej funkcie. Na obr. 7 je znázornený priebeh cross-validačnej funkcie a vyznačený bod $b_{0,2}^{CV} = 0.086$, v ktorom nadobúda minimum.



Obr. 7: Priebeh cross-validačnej funkcie a jej minimum

Voľba b pre odhady derivácií

Hľadanie optimálneho vyhladzovacieho parametra pre odhad derivácie je zložitejšie, pretože náš odhad derivácie nemáme s čím "porovnávať". Napriek tomu sa budeme snažiť o zobecnenie metód popísaných v predchádzajúcej stati.

Náš odhad derivácie tak budeme nakoniec porovnávať s odhadom derivácie získaným podľa (1.36).

Po tejto transformácii dát potom definujme funkcie

$$\mathsf{RSS}^{(\nu)}(b) := \frac{1}{n-\nu}\sum_{i=1}^{n-\nu}\left(Y_i^{(\nu)} - \widehat{f^{(\nu)}}(x_i^{(\nu)})\right)^2, \tag{2.57}$$

$$\mathsf{CV}^{(\nu)}(b) := \frac{1}{n-\nu}\sum_{i=1}^{n-\nu}\left(Y_i^{(\nu)} - \widehat{f^{(\nu)}_{-\{i,\ldots,i+\nu\}}}(x_i^{(\nu)})\right)^2, \tag{2.58}$$

$$\mathsf{R}^{(\nu)}(b) := \mathsf{RSS}^{(\nu)}(b) - \binom{2\nu}{\nu}\widehat{\sigma^2} + (-1)^\nu \frac{2\widehat{\sigma^2}}{nb^{2\nu+1}}K_\nu^{(\nu)}(0). \tag{2.59}$$

Odhady minimalizáciou (2.58), resp. (2.59) sú opäť konzistentné vzhľadom k IMSE v zmysle

$$\mathsf{E}\{\mathsf{R}^{(\nu)}(b)\} \to \mathsf{IMSE}(b) \text{ a } \mathsf{E}\{\mathsf{CV}^{(\nu)}(b)\} \to \mathsf{IMSE}(b) + \binom{2\nu}{\nu}\sigma^2, \tag{2.60}$$

odkiaľ vyplýva, že pre b zvolené na základe (2.58) alebo (2.59) platí $b = \mathcal{O}(n^{-\frac{1}{2k+1}})$ a b spĺňa predpoklady viet 2.14, 2.15 a 2.17.

Ďalšia metóda odhadu optimálneho vyhladzovacieho parametra pre odhad derivácií regresnej funkcie vychádza z odhadu (označíme ho $b_{0,k}$) parametra b, pre samotnú regresnú funkciu pri užití jadrovej funkcie rádu $(0,k)$.

Parametre jadrovej funkcie K_ν rádu (ν,k) označme postupne $V_{\nu,k}$ a $B_{\nu,k}$.

Zo vzťahov (2.15) a (2.13) dostávame hlavné členy IMSE:

$$\mathsf{IMSE}(b_{\nu,k}) \doteq \frac{\sigma^2 V_{\nu,k}}{nb_{\nu,k}^{2\nu+1}} + b_{\nu,k}^{2(k-\nu)}B_{\nu,k}^2\int\left(f^{(k)}(x)\right)^2dx. \tag{2.61}$$

Minimalizáciou tohoto vzťahu vzhľadom k b dostávame asymptoticky optimálny vyhladzovací parameter

$$b_{\nu,k}^* = \left(\frac{1}{n}\cdot\frac{2\nu+1}{2(k-\nu)}\cdot\frac{\sigma^2 V_{\nu,k}}{B_{\nu,k}^2\int\left(f^{(k)}(x)\right)^2dx}\right)^{\frac{1}{2k+1}}. \tag{2.62}$$

Porovnajme odhady $b_{\nu,k}^*$ a $b_{0,k}^*$:

$$\frac{b_{\nu,k}^*}{b_{0,k}^*} = \left((2\nu+1)\cdot\frac{k}{k-\nu}\cdot\frac{V_{\nu,k}}{V_{0,k}}\cdot\left(\frac{B_{0,k}}{B_{\nu,k}}\right)^2\right)^{\frac{1}{2k+1}} =: d_{\nu,k}^0. \tag{2.63}$$

Ak máme nejaký asymptoticky optimálny odhad $b_{0,k}^*$ (v rámci asymptotiky napríklad použime b_{CV} alebo b_R na základe (2.55), resp. (2.56)), potom dostaneme jednoducho

$$b_{\nu,k}^* = d_{\nu,k}^0 \cdot b_{0,k}^*. \tag{2.64}$$

Tento postup sa nazýva "*factor method*".

Táto konkrétna metóda je využiteľná iba pre ν párne (pretože *smooth optimum kernels* poznáme iba s rovnakou paritou), preto navrhnime analogický postup pre ν nepárne. Odhadnime $b^*_{1,k}$ na základe (2.58) alebo (2.59) pre odhad prvej derivácie a ďalej spočítajme (pre ν nepárne):

$$d^1_{\nu,k} := \frac{b^*_{\nu,k}}{b^*_{1,k}} = \left(\frac{2\nu+1}{3} \cdot \frac{k-1}{k-\nu} \cdot \frac{V_{\nu,k}}{V_{1,k}} \cdot \left(\frac{B_{1,k}}{B_{\nu,k}} \right)^2 \right)^{\frac{1}{2k+1}} . \tag{2.65}$$

V tab. 2 sú uvedené hodnoty $d^0_{\nu,k}$ a $d^1_{\nu,k}$ pre rôzne $(\nu, k, \mu = 1)$:

(ν, k)	$d^0_{\nu,k}$	(ν, k)	$d^1_{\nu,k}$
$(2, 4)$	0.8918992	$(3, 5)$	0.9177226
$(2, 6)$	0.9507335	$(3, 7)$	0.9603965
$(4, 6)$	0.8875622	$(5, 7)$	0.9068703
$(4, 8)$	0.9395030	$(5, 9)$	0.9485646

Tab. 2: Tabuľka niektorých hodnôt $d^0_{\nu,k}$ a $d^1_{\nu,k}$

Metódy hľadania optimálneho vyhladzovacieho parametra pre odhady derivácií sú popísané v [MuStSc87].

2.2 *Bootstrap* a konfidenčné množiny jadrových odhadov

Na zostrojenie konfidenčných množín pre naše odhady môžeme použiť centrálne limitné vety uvedené v stati 2.1.2. Tieto množiny sú však (najmä v prípade odhadov pre vyššie derivácie) približné, zbytočne "opatrne" veľké a naviac, vyžadujú odhady ďalších derivácií (vzorce vo vetách 2.14, 2.15, 2.17). Tieto skutočnosti budú ilustrované príkladom v kapitole 2.4.

V tejto časti nám k zostrojeniu presnejších intervalových odhadov, resp. konfidenčných množín pomôžu tzv. *resampling* metódy, konkrétne *bootstrap*. Pri riešení parametrických úloh sa používa napríklad pre odhady parametrov rozdelenia. Jeho podstata spočíva v opakovaných výberoch (napr. s opakovaním) z pozorovaní (napr. rovnomerne). V rámci týchto opakovaní sa napočítavajú vždy nové odhady parametra. Parameter sa potom chápe ako náhodná veličina s rozdelením, ktoré sa aproximuje rozdelením získaným na základe spomínaných opakovaných výberov podmienené pôvodným výberom. Odtiaľ sa spočítajú kritické hodnoty, resp. kvantily a za určitých predpokladov sme, ak možno uvažovať napr. normalitu, hotoví. Pri úlohách regresie sa namiesto opakovaných výberov z pozorovaní najčastejšie pracuje s odhadmi reziduí. Podobne budeme postupovať aj v našom regresnom modeli:

$$Y_i = f(x_i) + e_i; \quad e_i \sim i.i.d., \mathsf{E}e_i = 0, \mathsf{var}\ e_i = \sigma^2; \quad i = 1, \ldots, n. \tag{2.66}$$

Predpokladajme, že odhadujeme polohu nulového bodu ζ_ν funkcie $f^{(\nu)}$. Vieme, že odhady z kapitoly 2.1 vykazujú asymptotickú normalitu.

*Bootstrap*ový algoritmus pre nájdenie konfidenčných množín pre naše odhady, ktorý si teraz popíšeme, je miernou modifikáciou algoritmu uvedeného v [WaGa98].

Označme:

$$Z := \zeta_{n,\nu} - \zeta_\nu. \tag{2.67}$$

Pre nejaké $\alpha \in (0,1)$ definujme "kritickú normovanú šírku intervalu"

$$\varrho_\alpha = \inf\left\{\varrho; \mathsf{P}\left[\frac{Z}{\sqrt{\mathsf{var}\ Z}} \in (-\varrho, \varrho)\right] \geqq \alpha\right\}. \tag{2.68}$$

Takto sa dostávame k samotnému $(100\alpha)\%$ intervalu spoľahlivosti pre odhad nulového bodu

$$\mathcal{K}_\alpha = \left(\zeta_{n,\nu} - \varrho_\alpha\sqrt{\mathsf{var}\ Z}, \zeta_{n,\nu} + \varrho_\alpha\sqrt{\mathsf{var}\ Z}\right). \tag{2.69}$$

Ukážeme, že vhodná aproximácia rozdelenia náhodnej veličiny Z dáva aj vhodnú aproximáciu konfidenčnej množiny K_α.

Algoritmus 1 (*Bootstrap* pre Gasserove - Müllerove jadrové odhady). Postup je nasledovný:

(0) Na základe pozorovaní $(x_1, Y_1), \ldots, (x_n, Y_n)$ spočítame Gasserov - Müllerov odhad $\zeta_{n,\nu}$ voľbou vyhladzovacieho parametra $b := b_{\nu,k}^{CV}$ a jadrovej funkcie $K_{\nu,k,\mu}$ rádu (ν, k, μ).

(1) Na základe pozorovaní $(x_1, Y_1), \ldots, (x_n, Y_n)$ vypočítame odhad f_n^0 funkcie f s voľbou vyhladzovacieho parametra[12] b_0 a jadrovej funkcie rádu $K_{0,k-\nu,\mu+\nu}$ a odhad $\zeta_{n,\nu}^0$ bodu ζ_ν s voľbou vyhladzovacieho parametra b^0 a jadrovej funkcie $K_{\nu,k,\mu}$; potom odhad f_n funkcie f s voľbou vyhladzovacieho parametra $b_{0,k-\nu}^{CV}$ a jadrovej funkcie $K_{0,k-\nu,\mu+\nu}$.

(2) Spočítame "centrované" odhady reziduí

$$\widehat{e_i} = Y_i - f_n(x_i) - \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - f_n(x_j)\right). \tag{2.70}$$

(3) Z množiny $\{\widehat{e_i}; i = 1, \ldots, n\}$ vyberieme náhodne (rovnomerne a s opakovaním) *bootstrap*ové reziduá $\{e_i^*; i = 1, \ldots, n\}$.

(4) Vytvoríme *bootstrap*ové pozorovania (x_i, Y_i^*), pričom $Y_i^* = f_n^0(x_i) + e_i^*$, na základe ktorých spočítame nový *bootstrap*ový odhad $\zeta_{n,\nu}^*$ bodu ζ_ν s voľbou vyhladzovacieho parametra b a jadrovej funkcie $K_{\nu,k,\mu}$.

(5) Označme $Z^* := \zeta_{n,\nu}^* - \zeta_{n,\nu}^0$.

(6) Spočítajme odhad

$$\varrho_\alpha^* = \inf\left\{\varrho; \mathsf{P}\left[\frac{Z^*}{\sqrt{\mathsf{var}\ Z^*}} \in (-\varrho, \varrho) \mid e_1, \ldots, e_n\right] \geqq \alpha\right\}, \tag{2.71}$$

nakoniec spočítajme

$$\mathcal{K}_\alpha^* = \left(\zeta_{n,\nu} - \varrho_\alpha^*\sqrt{\mathsf{var}\ Z^*}, \zeta_{n,\nu} + \varrho_\alpha^*\sqrt{\mathsf{var}\ Z^*}\right). \tag{2.72}$$

Veta 2.21 (CLV pre *bootstrap*ovú veličinu). *Nech existujú všetky momenty náhodných veličín e_i. Nech $n^{\frac{1}{2k+3}}b^0 \to \infty$ a nech $k > 3\nu + 1$, $\mu \geqq 2$. Potom*

(i) Podmienené rozdelenie náhodnej veličiny $(nb^{2\nu+1})^{\frac{1}{2}}Z^*$ za $\{e_1, \ldots, e_n\}$ daných a rozdelenie náhodnej veličiny $(nb^{2\nu^+1})^{\frac{1}{2}}Z$ konvergujú pri $n \to \infty$ k takému istému normálnemu rozdeleniu;

(ii) $\mathsf{P}[Z \in \mathcal{K}_\alpha^*] \xrightarrow{n\to\infty} \alpha$.

[12] b_0 volíme podľa vety 2.21 a poznámky 2.22.

Dôkaz. Dôkaz sa nájde v [WaGa98] na str. 989-990. □

Poznámka 2.22. Podmienka $n^{\frac{1}{2k+3}}b^0 \to \infty$ nám hovorí, že počiatočný *bandwidth* b^0 je potrebné voliť rádu väčšieho ako $\asymp n^{\frac{-1}{2k+3}}$, teda aj väčšieho ako $\asymp n^{\frac{-1}{2k+1}}$ (takúto voľbu nazývame *oversmoothing*). Voľbou väčšieho *bandwidth*u zväčšíme vychýlenie *bootstrap*ového odhadu, ale zmenšíme rozptyl. Nutnosť voľby väčšieho vyhladzovacieho parametra v kroku (1) sa dá v krátkosti zjednodušene odôvodniť. Uvedomme si, že veličina Z^* je vlastne odhadom vychýlenia nášho pôvodného odhadu z kroku (0). Vychýlenie odhadu asymptoticky závisí na $f^{(k)}(\zeta_\nu)$, nuž a konzistentné odhadovanie funkcie k-tej derivácie takto zvyšuje rád konvergencie oproti "klasickému" $\mathcal{O}(n^{-\frac{1}{2k+1}})$. Viac o tejto preblematike pojednáva [HaMa91]. Navrhujeme voliť napríklad $b^0 := b^{CV}_{0,k+2}$, $b^{CV}_{0,k+3}$ alebo $b^{CV}_{1,k+2}$. Všetky voľby vyhladzovacieho parametra môžeme prevádzať aj na základe Riceovho kritéria (2.59). Takouto voľbou vlastne porovnávame priamo rozdiel reziduí medzi odhadom nadhladením (*oversmoothing*om) a odhadom voľbou optimálneho parametra (konzistentného vzhľadom k IMSE). Kombinácia metód je takisto teoreticky prípustná.

Poznámka 2.23. Podmienené rozdelenie sa aproximuje empirickým rozdelením realizácií náhodnej veličiny Z^* získaných na základe dostatočne veľkého počtu opakovaných simulácií krokov (3)-(5) algoritmu 1.

Poznámka 2.24. *Bootstrap* sa používa aj pre testovanie hypotéz.

V kapitole 2.4 porovnáme interval spoľahlivosti založený na metóde *bootstrap* s intervalom spoľahlivosti založeným na centrálnej limitnej vete (kap. 2.1.2).

2.3 Jadrové odhady s korelovanými chybami

V tejto kapitole popíšeme fungovanie jadrových odhadov pre model s korelovanými chybami. Budeme sa snažiť o analógie predchádzajúcich viet, resp. metód. Často budeme potrebovať opakované pozorovania. Vychádzame z článku [Am85]. Analógie odvodíme iba pre určitý tvar kovariančnej štruktúry.

Pre tento prípad vychádzajme z modelu:

$$Y_j(x_i) = f(x_i) + e_j(x_i); \quad i = 1, \ldots, n; \quad j = 1, \ldots, m \tag{2.73}$$

a predpokladajme, že platia predpoklady 2.25 a 2.26.

Predpoklady 2.25. Požadujeme:

(A1) $f \in \mathcal{C}^2([0,1])$,

(A2) $x_i \in [0,1]$ sú asymptoticky ekvidistantné v zmysle predpokladov 2.2.

Predpoklady 2.26. O rozdelení chýb budeme postupne predpokladať:

(A3) $\mathsf{E}e_i(\cdot) \equiv 0$ a kovariančná štruktúra má tvar

$$\mathsf{cov}\left(e_j(x_i), e_k(x_l)\right) = \begin{cases} \sigma^2\gamma_\theta\left(T_\lambda(x_i) - T_\lambda(x_l)\right) & j = k \\ 0 & j \neq k \end{cases}, \tag{2.74}$$

(A4) $\sigma > 0$, γ_θ a T_λ sú dané parametrické funkcie,

(A5) $\gamma_\theta(\cdot)$ je párna, lipschitzovská a $\gamma'_\theta(0_+) < 0$,

(A6) $T_\lambda(x)$ je rýdzomonotónna v x a diferencovateľná v x pre všetky $x \neq 0$,

(A7) $T_\lambda(x)$ je hölderovská v 0, čiže splňuje

$$\exists_{\substack{\delta\in(0,1] \\ H_\delta\in\mathbb{R}}} : \lim_{\Delta\to 0_+} \sup_{x\in[0,1]} \frac{T_\lambda(x+\Delta) - T_\lambda(x)}{\Delta^\delta} = H_\delta, \tag{2.75}$$

(A8) $\exists_{\substack{\tau\geqq 0 \\ K\in\mathbb{R}}} \forall_{i,l;j} : \mathsf{E}|e_j(x_i)e_j(x_l)|^{4+\tau} \leqq K$,

(A9) γ_θ a T_λ sú dvakrát diferencovateľné vzhľadom k θ a λ v zmienenom poradí.

Poznámka 2.27. Vzhľadom k tomu, že $\gamma_\theta(\cdot)$ je párna, bez újmy na obecnosti môžeme predpokladať, že $T_\lambda(\cdot)$ je rastúca.

K jednoznačnému určeniu funkcií γ_θ a T_λ môžeme tiež stanoviť $\gamma_\theta(0) = 1$, $T_\lambda(1) = 0$ a $T'_\lambda(1) = 1$. Všimnime si tiež, že ak $T_\lambda(x) = x - 1$, proces bude stacionárny.

θ a λ môžu byť aj vektorové parametre.

Príklad 2.28 (nestacionárnej kovariančnej štruktúry skupiny nezávislých náhodných procesov). Núñez-Antón a Woodworth navrhli (1994):

$$\mathsf{cov}\left(e_j(x_i), e_k(x_l)\right) = \begin{cases} \sigma^2 \rho^{|x_i^\lambda - x_l^\lambda|/\lambda} & j = k \\ 0 & j \neq k \end{cases}, \tag{2.76}$$

kde $\sigma > 0$, $\rho \in (0,1)$ a $\lambda > 0$ sú parametre.

V tomto prípade je $\theta = \rho$, $\gamma_\rho(t) = \rho^{|t|}$ a $T_\lambda(x) = \frac{x^\lambda - 1}{\lambda}$ je Boxova-Coxova transformácia. Je ľahké ukázať, že predpoklady 2.26 sú splnené a je jasné, že pre $\lambda = 1$ ide o stacionárny proces.

Prejdime teraz k samotnému jadrovému odhadu regresnej funkcie f, ide o prirodzené zobecnenie vzťahu (2.7):

$$\hat{f}(x) := f_n(x) = \frac{1}{b} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K\left(\frac{x-t}{b}\right) dt \cdot Y_\bullet(x_i), \tag{2.77}$$

kde

$$Y_\bullet(x_i) = \frac{1}{m} \sum_{j=1}^{m} Y_j(x_i). \tag{2.78}$$

2.3.1 Vyhladzovací parameter

Podobne ako v kapitole 2.1.5, aj pre model s korelovanými chybami sa budeme snažiť nájsť optimálny vyhladzovací parameter. Uvidíme, že nám bude stačiť zobecniť Riceovo kritérium (2.56) na (2.96). Postupovať budeme dôslednejšie a podrobnejšie ako v kapitole 2.1.5.

Veta 2.29 (MSE jadrového odhadu). *Nech platia predpoklady (A1)-(A7). Nech $x \in (0,1)$ a $f''(t) \neq 0$. Nech jadrová funkcia K je rádu $(0,2)$ s nosičom [-1,1], potom pre $n, m \to \infty$, $h \to 0$ platí*

$$\mathsf{MSE}\left(f_n(x)\right) \approx \frac{\sigma^2}{m}\left(1 + b\gamma_\theta'(0_+)T_\lambda'(x)C\right) + \frac{b^4}{4}\left(f''(x)\right)^2 B^2 + \\ + o(b^4) + \mathcal{O}(n^{-2} + (mn^\delta)^{-1} + b^2 n^{-1}), \tag{2.79}$$

ak označíme $B := 2B_2$ a $C := \int\int |r-t| K(r)K(t) dr dt$.

Dôkaz. Budeme hľadať asymptotické vyjadrenie pre $\mathsf{var}\ f_n(x)$ a $\mathsf{bias} f_n(x)$, pretože vieme, že $\mathsf{MSE} = \mathsf{var}\ + \mathsf{bias}^2$.

$$f_n(x) - \mathsf{E} f_n(x) = \frac{1}{b} \sum_{i=1}^{n} (Y_\bullet(x_i) - f(x_i)) \int_{s_{i-1}}^{s_i} K\left(\frac{x-t}{b}\right) dt. \tag{2.80}$$

Taylorovým rovojom funkcie $\mathsf{E}f_n(x) - f(x)$ do 2. rádu dostávame

$$\mathsf{bias}(f_n(x)) = \frac{b^2}{2} B f''(x) + o(b^2) + \mathcal{O}\left(\frac{1}{n}\right). \tag{2.81}$$

Zo vzorca (2.80) ďalej dostávame:

$$\begin{aligned} \mathsf{var}\ f_n(x) &= \mathsf{E}(f_n(x) - \mathsf{E}f_n(x))^2 = & (2.82)\\ &= \frac{1}{b^2} \sum_{i,l} \mathsf{E}\left\{e_\bullet(x_i) e_\bullet(x_l)\right\} \int_{s_{i-1}}^{s_i} \int_{s_{l-1}}^{s_l} K\left(\frac{x-r}{b}\right) K\left(\frac{x-t}{b}\right) dr dt \end{aligned}$$

$$\text{a} \qquad \mathsf{E}\left\{e_\bullet(x_i) e_\bullet(x_l)\right\} = \frac{\sigma^2}{m} \gamma_\theta(T_\lambda(x_i) - T_\lambda(x_l)). \tag{2.83}$$

V tomto okamihu budeme potrebovať k aproximácii rozptylu určitú aproximáciu kovariancie, o ktorej hovorí nasledujúca lemma.

Lemma 2.30 (Aproximácia kovariancie). *Ak $r \in [s_{i-1}, s_i]$ a súčasne $t \in [s_{l-1}, s_l]$, potom*

$$\gamma_\theta(T_\lambda(x_i) - T_\lambda(x_l)) - \gamma_\theta(T_\lambda(r) - T_\lambda(t)) = \mathcal{O}(n^{-\delta}). \tag{2.84}$$

Dôkaz lemmy 2.30. Použijeme lipschitzovskosť funkcie γ_θ podľa predpokladu (A5), hölderovskosť funkcie T_λ podľa predpokladu (A7), a tiež využijeme asymptotickú ekvidistanciu.

$$\begin{aligned} |\gamma_\theta(T_\lambda(x_i) - T_\lambda(x_l)) - \gamma_\theta(T_\lambda(r) - T_\lambda(t))| &\leqq \textit{(lipschitzosvskosť)}\\ \leqq C|(T_\lambda(x_i) - T_\lambda(x_l)) - (T_\lambda(r) - T_\lambda(t))| &= \\ = C|(T_\lambda(x_i) - T_\lambda(r)) - (T_\lambda(x_l) - T_\lambda(t))| &\leqq \textit{(monotónia } T_\lambda(\cdot) \textit{ a } \triangle)\\ \leqq C^* \left[(T_\lambda(s_i) - T_\lambda(s_{i-1})) + (T_\lambda(s_l) - T_\lambda(s_{l-1}))\right] &\leqq C^{**} \sup_\kappa \left[T_\lambda(s_\kappa) - T_\lambda(s_{\kappa-1})\right]. \end{aligned}$$

Vďaka asymptotickej ekvidistancii a hölderovskosti existuje

$$\lim_{n\to\infty} \sup_\kappa n^{-\delta} \left[T_\lambda(s_\kappa) - T_\lambda(s_{\kappa-1})\right].$$

Koniec dôkazu lemmy 2.30. □

Použitím rovností (2.82) a (2.83), lemmy 2.30 a subst. $u := \frac{x-r}{b}$ a $v := \frac{x-t}{b}$ dostávame

$$\mathsf{var}\ f_n(x) \approx \frac{\sigma^2}{m} \cdot J + \mathcal{O}(m^{-1} n^{-\delta}), \tag{2.85}$$

$$\text{pritom} \quad J = \int_{-1}^{1} \int_{-1}^{1} \rho_b(u,v) K(u) K(v) du dv, \tag{2.86}$$

$$\text{kde} \quad \rho_b(u,v) = \gamma_\theta(T_\lambda(x - bu) - T_\lambda(x - bv)). \tag{2.87}$$

Teraz stačí ukázať, že

$$J = 1 + b\gamma'_\theta(0_+)T'_\lambda(x)C + o(b) \tag{2.88}$$

a tvrdenie vety bude dokázané. Všimnime si, že $\rho_0(u,v) = 1$. Zapíšme

$$J = 1 + b \int\int \frac{\rho_b(u,v) - 1}{b} K(u)K(v)dudv, \tag{2.89}$$

pretože $\int K = 1$, všimnime si, že vzhľadom k definícii derivácie, vete o derivácii zloženej funkcie a predpokladu (A5) platí

$$\frac{\rho_b(u,v) - 1}{b} \xrightarrow{b\to 0} \rho'_{0_+}(u,v) = \gamma'_\theta(0_+)T'_\lambda(x)|v-u|. \tag{2.90}$$

Vďaka predpokladom (A5) a (A7) môžeme vo vyjadrení (2.89) podľa Lebesgueovej vety zameniť limitu a integrál a sme hotoví.
Dôkaz je aj v článku [Fe97], str. 418-420. □

Poznámka 2.31. Pre $m, n \to \infty$, $b \to 0$ nepotrebujeme k bodovej konzistencii "klasický" požiadavok $nb \to \infty$. Pre m pevné a $nb \to \infty$ platí $\mathsf{MSE}(f_n(x)) \to \frac{\sigma^2}{m}$.

Ďalej označme $\mathsf{MSE}(x,b) := \mathsf{MSE}(f_n(x;b))$.

Veta 2.32 (Lokálny optimálny *bandwidth* pri korelovaných chybách). *Ak sú splnené predpoklady vety 2.29, potom pre* $\frac{n^\delta}{m} = \mathcal{O}(1)$ *(*δ *splňujúca predpoklad (A7)) platí*

$$\mathsf{MSE}(x, b^*_m(x)) \leqq \mathsf{MSE}(b_{n,m}(x)), \tag{2.91}$$

pričom

$$b^*_m(x) = \left[-\frac{\sigma^2 C\gamma'_\theta(0_+)T'_\lambda(x)}{B^2[f''(x)]^2}\right]^{\frac{1}{3}} m^{-\frac{1}{3}} \text{ je asympt. optimálny} \tag{2.92}$$

$$a \quad b_{n,m}(x) = \left[\frac{\sigma^2 V}{B^2[f''(x)]^2}\right]^{\frac{1}{5}} (mn)^{-\frac{1}{5}} \text{ je as. opt. pre i.i.d. chyby.} \tag{2.93}$$

Nerovnosť (2.91) *je ostrá práve vtedy, keď chyby nie sú nezávislé. Pre stacionárne chyby (nie nezávislé) platí*

$$b^{**}_m(x) = \left[-\frac{\sigma^2 C\gamma'_\theta(0_+)}{B^2[f''(x)]^2}\right]^{\frac{1}{3}} m^{-\frac{1}{3}} \quad a \quad \mathsf{MSE}(x, b^{**}_m(x)) < \mathsf{MSE}(x, b_{n,m}(x)). \tag{2.94}$$

Ďalej platí, že $\mathsf{MSE}(x, b^*_m(x)) < \mathsf{MSE}(x, b^{**}_m(x))$ *práve vtedy, keď chyby nie sú stacionárne.*

Poznámka 2.33. Nerovnosti chápeme asymptoticky, v zmysle

$$g_{n,m}(t) < h_{n,m}(t) \Leftrightarrow \exists_{n_0,m_0}\forall_{n>n_0,m>m_0}\forall_t :\ g_{n,m}(t) < h_{n,m}(t).$$

Dôkaz. Použitím vzťahu (2.79) a položením derivácie jeho hlavných členov rovnej nule. □

Ak si uvedomíme, že $\mathsf{IMSE}(b) = \mathsf{MISE}(b) = \int_0^1 \mathsf{MSE}(x,b)dx$, potom ľahko dostávame aj optimálny vyhladzovací parameter vzhľadom k IMSE:

$$b_m^* = \left[-\frac{\sigma^2 C \gamma_\theta'(0_+)\left(T_\lambda(1) - T_\lambda(0)\right)}{B^2 \int_0^1 [f''(x)]^2 dx} \right]^{\frac{1}{3}} m^{-\frac{1}{3}} = \left[\frac{\sigma^2 C \gamma_\theta'(0_+)\left(T_\lambda(0)\right)}{B^2 \int_0^1 [f''(x)]^2 dx} \right]^{\frac{1}{3}} m^{-\frac{1}{3}}. \tag{2.95}$$

Vzhľadom k tomu, že funkcia f je neznáma, potrebujeme pre voľbu b kritérium, ktoré na f nebude závisieť. Jednou z možností je zobecnenie Riceovho kritéria $\mathsf{R}(b)$ zo vzorca (2.56):

$$\mathsf{R}(b) := \mathsf{RSS}(b) - \frac{\widehat{\sigma^2}}{m} + \frac{2\widehat{\sigma^2}}{nmb} \sum_{i,l=1}^{n} \gamma_{\hat\theta}(T_{\hat\lambda}(x_i) - T_{\hat\lambda}(x_l)) \int_{s_{l-1}}^{s_l} K\left(\frac{x_i - t}{b}\right) dt, \tag{2.96}$$

kde

$$\mathsf{RSS}(b) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_\bullet(x_i) - f_n(x_i)\right)^2 \tag{2.97}$$

a $(\widehat{\sigma^2}, \hat\theta, \hat\lambda)$ je konzistentný odhad vektoru $(\sigma^2, \theta, \lambda)$.

Veta 2.34 (Globálny optimálny *bandwidth*). *Ak sú splnené predpoklady vety 2.32 a je splnený predpoklad (A8), potom odhad b_m^R minimalizujúci Riceovo kritérium $\mathsf{R}(b)$ zo vzťahu (2.96) je konzistentný vzhľadom k IMSE v zmysle*

$$\arg\min \mathsf{R}(b) \to \arg\min \mathsf{IMSE}(b). \tag{2.98}$$

Dôkaz. Je dosť technický, nájde sa v článku [Fe97], str.420-422, prebieha opakovanou aplikáciou Minkowského nerovnosti "na predpoklad (A8)". □

Vo všeobecnosti o neparametrickej regresii s korelovanými chybami pojednáva napr. článok [OpWaYa01], popísaný je tu aj problém zobecnenia cross-validačnej metódy (stať 3.1., str. 138-140).

2.3.2 $\widehat{\mathsf{cov}}$

Ku vhodnému nastaveniu vyhladzovacieho parametra pre náš odhad potrebujeme ešte konzistentný odhad parametrov kovariančnej funkcie. Postup je jednoduchý a vychádza z metódy najmenších štvorcov.

Definujme empirickú kovariančnú funkciu

$$\widehat{\mathsf{cov}_{i,l}} := \frac{1}{m-1} \sum_{j=1}^{m} \left(Y_j(x_i) - Y_\bullet(x_i)\right)\left(Y_j(x_l) - Y_\bullet(x_l)\right), \quad i,l = 1,\dots n. \tag{2.99}$$

Vzhľadom k tomu, že $\mathsf{E}\widehat{\mathsf{cov}}_{i,l} = \sigma^2\gamma_\theta(T_\lambda(x_i) - T_\lambda(x_l))$, môžeme za predpokladu (A9) dostať $\sqrt{n^2}$-*konzistentný* odhad kovariančnej funkcie minimalizáciou kritéria

$$\mathcal{Q}_n(\sigma^2, \theta, \lambda) = \frac{1}{n^2}\sum_{i,l=1}^{n}\left\{\widehat{\mathsf{cov}_{i,l}} - \sigma^2\gamma_\theta(T_\lambda(x_i) - T_\lambda(x_l))\right\}^2, \tag{2.100}$$

ak naviac platí, že empirická kovariančná funkcia má ohraničenú množinu vlastných čísel. Viac o tejto úlohe možno nájsť napríklad v [Am85].

2.3.3 Známa variančná matica

Všimnime si, že uvažovaná kovariančná štruktúra v predchádzajúcej stati vlastne predpokladá homoskedasticitu v rámci j-teho výberu a tiež, variančnú maticu vektorov chýb $\boldsymbol{e}_j = (e_j(x_1), \ldots, e_j(x_n))^T$ nepozná a odhaduje, a na druhej strane jej predpisuje určitý tvar. Ako sa zmení náš odhad vyhladzovacieho parametra, ak variančnú maticu vektoru chýb poznať budeme?

Veta 2.35 (Riceovo kritérium pre známu variančnú maticu). *Predpokladajme, že*

$$Y_j(x_i) = f(x_i) + e_{j_i}, \tag{2.101}$$

$\boldsymbol{e}_j := (e_{j_1}, \ldots, e_{j_n})^T$, $\mathsf{E}\boldsymbol{e}_j = \boldsymbol{0}$, $\mathsf{var}\ \boldsymbol{e}_j = \sigma^2\boldsymbol{W}^{-1}$, $\boldsymbol{W} > 0$ *známa. Vyhladzovací parameter pre odhad regresnej funkcie, ktorého voľba je založená na Riceovom kritériu* (2.96)*, v ktorom odhad* $\gamma_{\hat\theta}(T_{\hat\lambda}(x_i) - T_{\hat\lambda}(x_l))$ *nahradíme známou skutočnou hodnotou* $v_{i,l}$*, pričom* $\boldsymbol{W}^{-1} = \boldsymbol{V} = (v_{i,l})$ *a odhad parametra* σ^2 *založíme na metóde najmenších štvorcov:*

$$\widehat{\sigma^2} = \arg\min_{\sigma^2}\frac{1}{n^2}\sum_{i,l}\left\{\widehat{\mathsf{cov}_{i,l}} - \sigma^2 v_{i,l}\right\}^2, \tag{2.102}$$

je konzistentný vzhľadom k IMSE *v zmysle* (2.60)*.*

Dôkaz. Je založený na rovnosti:

$$\mathsf{E}[\mathsf{RSS}(b)] = \frac{\sigma^2}{m} + \mathsf{IMSE}(b) - \frac{2\sigma^2}{mnb}\sum_{i,l=1}^{n} v_{i,l}\int_{s_{l-1}}^{s_l} K\left(\frac{x_i - t}{b}\right)dt. \tag{2.103}$$

□

2.3.4 Intervalové odhady nulových bodov

K určeniu intervalových odhadov môžeme použiť tzv. *wild bootstrap*, ako je spomenuté v [Ya03] na str. 161 a str. 157. Oproti klasickému *bootstrap*u sa tu rovnomerne

náhodne vyberá s opakovaním z $\{\widehat{e_{\bullet_i}} \cdot w_i\}$, kde w_i je náhodná veličina s diskrétnym rozdelením

$$\mathsf{P}\left[w_i = \frac{1-\sqrt{5}}{2}\right] = \frac{5+\sqrt{5}}{10}, \quad \mathsf{P}\left[w_i = \frac{1+\sqrt{5}}{2}\right] = \frac{5-\sqrt{5}}{10}, \tag{2.104}$$

pričom používame "necentrované" odhady reziduí $\widehat{e_{\bullet_i}}$, takže

$$\widehat{e_{\bullet_i}} = Y_\bullet(x_i) - f_n(x_i). \tag{2.105}$$

Poznámka 2.36. Platí $\mathsf{E}w_i = 0$, $\mathsf{E}(w_i^2) = 1$ a $\mathsf{E}(w_i^3) = 1$.

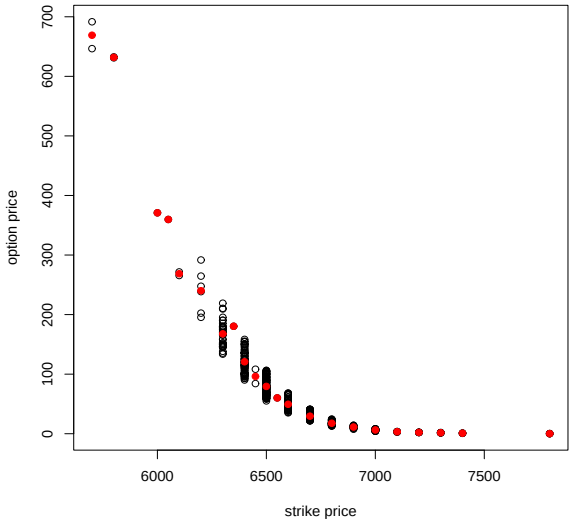
Poznámka 2.37. Pozor však na to, že *wild bootstrap* je použiteľný iba pre jednorozmerný odhad.

2.4 Aplikácia jadrových odhadov na reálne dáta

Nemecký akciový index DAX (Deutsche Aktienindex) je spoločným indexom 30-tich vybraných nemeckých cenných papierov obchodovaných na frankfurtskej burze (Deutsche Börse AG).

Našimi vstupnými dátami sú ceny 561 opcií typu *call* na DAX z 1.1.2001. Nezávisle premennou je cena akcie (*strike price*), závisle premennou bude cena opcie (*call option price*) upravená o diskontný faktor. Z ekonometrickej teórie ([HaHl05]) vyplýva, že druhá derivácia *option price* ako funkcie *strike price* je hustotou nejakého rozdelenia. Pre praktické aplikácie je vhodné odhadovať modus tohoto rozdelenia, teda maximum druhej derivácie, ktorého polohu odhadneme jadrovým odhadom. Intervalový odhad polohy maxima budeme konštruovať analógiou *bootstrap*ového algoritmu 1 (pre odhad polohy maxima). Tento intervalový odhad porovnáme s odhadom na základe CLV (pre odhad polohy maxima; kap. 2.1.2, veta 2.15).

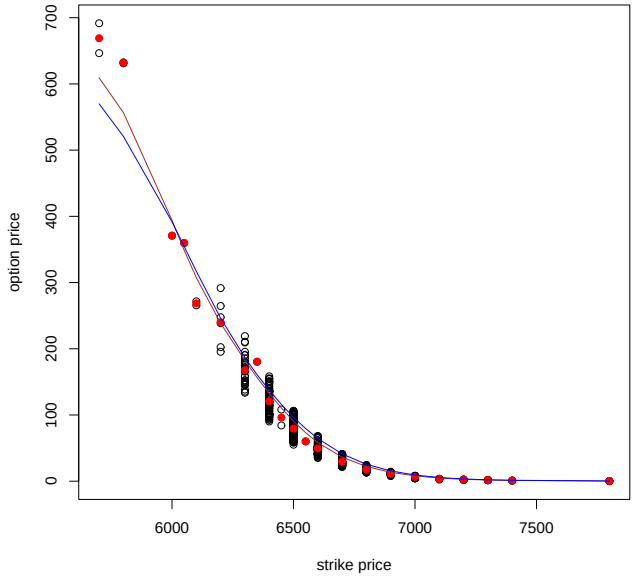
K dispozícii máme opakované pozorovania pre 22 rôznych cien *strike price*, pre ktoré spočítame priemery z daných opakovaných pozorovaní (rôzne *option price*). Na obr. 8 sú znázornené vstupné dáta. Červenou farbou sú na obr. 8 vyznačené priemery pre dané opakované pozorovania.



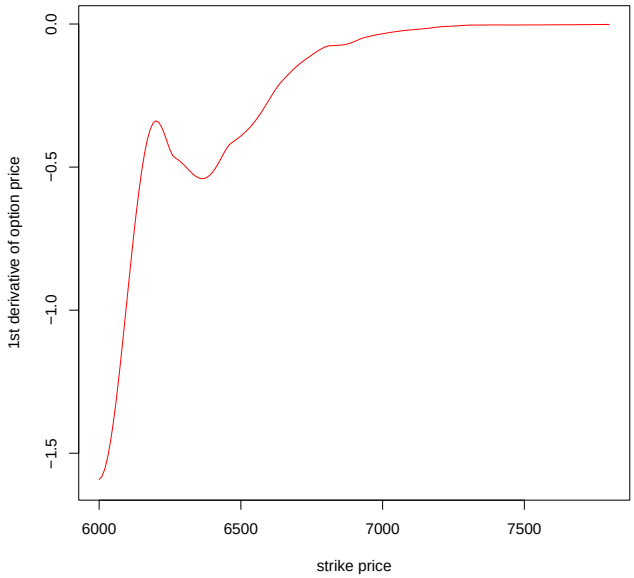
Obr. 8: Vstupné dáta (*strike price*, *option price*) a priemery opakovaných pozorovaní

Na obr. 9 znázorníme odhad priebehu regresnej funkcie - závislosti *option price* na *strike price*. Obr. 9 zároveň porovnáva odhady priebehov funkcií pre rôzne hodnoty vyhladzovacieho parametra z kroku (1) algoritmu 1, teda pre nadhladenie (modrá krivka) a pre optimálnu voľbu parametra (hnedá). Obr. 10 vykresľuje

odhad priebehu prvej derivácie regresnej funkcie, ktorá by mala byť neklesajúca (pretože jej derivácia má byť hustota). Toto obmedzenie ale v tomto príklade do riešenia nepremietame. Parametre zvolené pre jednotlivé jadrové odhady sú uvedené v tab. 3.

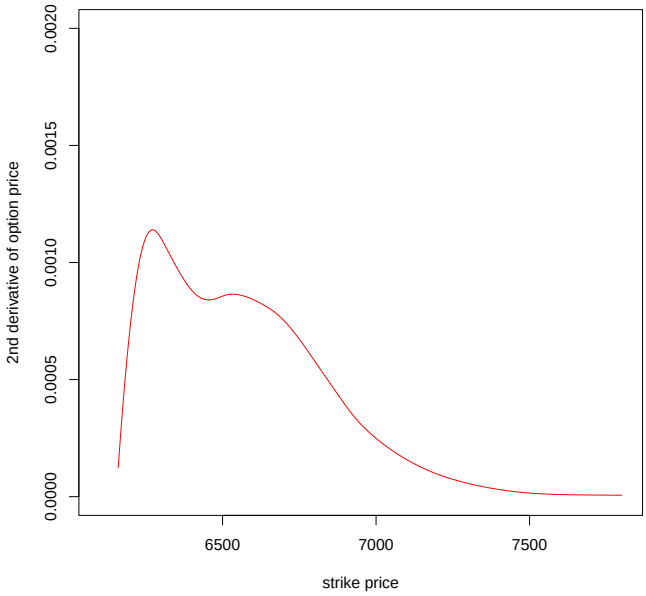


Obr. 9: Odhad *option price* ako funkcie *strike price*



Obr. 10: Odhad prvej derivácie *option price* ako funkcie *strike price*

Obrázky 9 a 10 sú iba ilustratívne - kľúčovú úlohu bude pre nás hrať odhad druhej derivácie, ktorý je znázornený na obr. 11.



Obr. 11: Odhad druhej derivácie *option price* ako funkcie *strike price*

Bodový odhad maxima vychádza $\theta_{22,2} = 6272$.

V tabuľke 3 uvedieme parametre pre jadrové odhady funkcií z obrázkov 9, 10 a 11, ktorých hodnoty sú získané metódou *cross-validation*. Posledný stĺpec tabuľky 3 zároveň pripomína, v ktorom kroku algoritmu 1 daný parameter používame.

Parameter	Hodnota	(ν, k, μ)	Obrázok	Krok algoritmu
b	608	$(2, 4, 2)$	11	(0) a (4)
b^0	565	$(1, 7, 1)$	9 (modrá), 10	(1)
b^{CV}	429	$(0, 2, 4)$	9 (hnedá)	(1)

Tab. 3: Parametre pre jadrové odhady (obr. 9, 10 a 11)

Spočítajme ešte intervalový odhad polohy maxima druhej derivácie založený na *bootstrap*ovom algoritme (veta 2.21) a porovnajme ho s intervalovým odhadom založenom na základe CLV (veta 2.15). Realizácie náhodnej veličiny Z^* boli simulované $200\times$.

Výsledky sú nasledovné:

(A) Pre *bootstrap*:

Smerodatná odchýlka *bootstrap*ovej veličiny Z^*:
$\sqrt{\mathsf{var}\ Z^*}$ =131.1006.

95% intervalový odhad polohy maxima druhej derivácie:
x.left=6015 x.est=6272 x.right=6529

(B) Pre CLV: Odhad $f_{22,4}(6272)$ hodnoty $f^{(4)}(\theta_2)$ sme počítali na základe jadrovej funkcie rádu $(4, 8, 1)$, vyhladzovací parameter určila metóda *cross-validation*. Spočítali sme nasledové odhady:

$\widehat{\sigma^2}$ =589.739
$V'_{4,8,1}$ =207516684
$b_{4,8,1}$ =725
$f_{22,4}(6272)$ =3.189872e-08

95% intervalový odhad polohy maxima druhej derivácie:
x.left=5825 x.est=6272 x.right=6719

Vidíme, že *bootstrap* dáva naozaj lepšie výsledky ako CLV.

Poznámka 2.38. Dáta k príkladu autorovi poskytol Mgr. Zdeněk Hlávka, Ph.D.

3 Prostredie R

V tejto kapitole budú predstavené niektoré podstatné časti zdrojového kódu a tiež spôsoby volania naprogramovaných funkcií. Kompletné komentované zdrojové kódy sú sprístupnené na mojej webovej stránke:

`http://artax.karlin.mff.cuni.cz/~jurij1am/dp`

3.1 Príklad 1.21

Odhady parametrov lineárneho modelu počíta (okrem iného) funkcia `lm()` a vyvolajú sa príkazom `coef(lm())`. Uvedený príklad bol počítaný pomocou funkcie `inter0()` zo súboru `linear.R`, časť kódu je prevzatá z [Zv04].

3.2 Príklad 1.22

Príklad bol počítaný funkciou `inter()` zo súboru `linear.R` a funkciou `predi()` zo súboru `pred.R`.

3.3 Príklad 1.25

Odhady parametrov modelu nelineárnej regresie počíta funkcia `nls()`, pristupuje sa k nim príkazom `coef(nls())`. K tomuto príkladu sa vzťahuje funkcia `intern()` zo súboru `nonlinear.R` a funkcia `transf()` zo súboru `trans.R`. Uvádzame skrátenú verziu funkcie `intern()`.

```
intern<-function(betamin=c(-1,-1),betamax=c(1,1),
                 plr=FALSE,x,y,N=500,eps=0.001)
# pocita konfidencnu mnozinu pre odhad parametrov
# na zaklade testu pomerom vierohodnosti
# pocita priblizny intervalovy odhad (xl2,xr2)
# betamin, betamax : "najsirsie mozne" hranice parametrov
# plr : vykreslit konfidencnu mnozinu LR testu
# x,y : data
# N : presnost - pocet bodov vycislovania na int. (betamin,betamax)
# eps : "hrubka" hranice konfidencnej mnoziny
{
  # parametricky predpis funkcie
  f<-function(a,c,iks)
  {
    f<-a*sin(c*pi*iks)
  }
  # odhad parametrov
  b<-coef(nls(y~f(a,c,x),data=list(c(x,y)),start=c(a=2,c=1)))
  xl2<-1/2/b[2]
  xr2<-xl2
  n<-length(x)
```

```
  k<-length(b)
  if (plr==TRUE)
  {
    plot(b[1],b[2],xlim=c(betamin[1],betamax[1]),
                   ylim=c(betamin[2],betamax[2]),pch=16)
  }
  argxl2<-b
  argxr2<-b
  s2<-sum((f(b[1],b[2],x)-y)^2)/(n-k)
  ef<-qf(0.95,k,n-k)
  # test pomerom vierohodnosti
  st<-sum((f(b[1],b[2],x)-y)^2)
  for (i in 1:(N+1))
  {
    for (j in 1:(N+1))
    {
      beta<-c(betamin[1]+(i-1)*(betamax[1]-betamin[1])/N,
              betamin[2]+(j-1)*(betamax[2]-betamin[2])/N)
      sb<-sum((f(beta[1],beta[2],x)-y)^2)
      if (sb <= st*(1+k*ef/(n-k)))
      {
        if (plr==TRUE)
        {
          if (abs(sb - st*(1+k*ef/(n-k))) <eps)
          {
            points(beta[1],beta[2],pch=16)
          }
        }
        if (1/(2*beta[2])<xl2)
        {
          xl2<-1/(2*beta[2])
          argxl2<-beta
        }
        if (1/(2*beta[2])>xr2)
        {
          xr2<-1/(2*beta[2])
          argxr2<-beta
        }
      }
    }
  }
  if (plr==TRUE)
  {
    points(argxl2[1],argxl2[2],col="yellow",pch=19)
    points(argxr2[1],argxr2[2],col="yellow",pch=19)
  }
  points(b[1],b[2],col="green",pch=19)
  est<-1/2/b[2]
  return(s2,xl2,xr2,argxl2,argxl2,est)
}
```

Ukážeme si ešte ako sa spočíta matica $\boldsymbol{F}(\boldsymbol{t})$, ktorá je potrebná pre výpočet intervalu (1.34).

```
# pocitanie matice F
F<-matrix(0,n,k)
df<-deriv(~a*sin(c*pi*iks),c("a","c","iks"),func=TRUE)
for (i in 1:n)
{
  for (j in 1:k) F[i,j]<-attr(df(b[1],b[2],x[i]),"gradient")[j]
}
```

3.4 Príklady 2.12, 2.20, tabuľky 1 a 2, kapitola 2.4

Všetky fukcie týkajúce sa jadrovej regresie sú v súbore `gamureg.R`. Pretože ich je viac, uvedieme ich zoznam, aj s krátkym popisom výstupu. Vstup je popísaný priamo v zdrojovom kóde. Funkcie potrebujú knižnicu `MASS`.

Funkcia	Výstup
`optkern`	Koeficienty jadrového polynómu
`integ`	Určitý integrál z polynomiálnej funkcie
`gamur`	Jadrový odhad funkcie / derivácie v jednom bode
`gamure`	Jadrový odhad funkcie / derivácie vo viacerých bodoch intervalu
`cv`	Funkčné hodnoty $\mathsf{CV}(b)$ vo viacerých bodoch intervalu
`squa`	Koeficienty druhej mocniny jadrového polynómu
`der`	Koeficienty prvej derivácie jadrového polynómu
`varian`	Hodnoty V, V' a B_k
`dnk`	Hodnoty $d^0_{\nu,k}$ alebo $d^1_{\nu,k}$
`transf`	Transfrmované dáta pre odhad derivácie
`me`	Vektor stredných hodnôt opakovaných pozorovaní
`boot`	Realizácie *bootstrap*ovej veličiny Z^*

Tab. 4: Zoznam naprogramovaných funkcií k jadrovej regresii

Pomocou funkcií `optkern`, `varian`, `squa`, `der` a `dnk` sú vyplnené tab. 1 a 2.

Pre názornosť si ukážeme malú modifikáciu funkcie `gamur`, ktorou sme počítali príklad 2.12.

```
gamur<-function(x,X,Y,nu=0,k=2,mu=1,b)
# Gasserov - Mullerov odhad nu-tej derivacie regresnej funkcie v bode x
# x : bod, v ktorom odhadujeme f^(nu)(x)
# X, Y : data
# nu, k, mu : rady jadrovej funkcie
# b : bandwidth
{
  # spocitame koeficienty jadroveho polynomu
  cK<-optkern(nu,k,mu)
  n<-length(X)
```

```
  X<-c(X[1],X,X[n])
  w<-numeric(n)
  for(i in 1:n)
  {
    w[i]<-integ((x-((1/2)*(X[i+1]+X[i])))/b,(x-((1/2)*(X[i+2]+X[i+1])))/b,cK)
  }
  fx<-(sum(w*Y))/(b^nu)
  # vystupom je odhad fx
  fx
}
```

Parameter b sme volili v príklade 2.12 metódou cross-validácie. V príklade 2.20 sme si ukázali tiež priebeh funckie $\mathsf{CV}(b)$. Uvádzame preto kratšiu verziu funkcie `cv`.

```
cv<-function(X,Y,nu=0,k=2,mu=1,left=0,right=1,pt=50)
# pocita "priebeh" funkcie CV
# X,Y : data
# nu,k,mu : rady jadrovej funkcie
# left, right hranice pre bandwidth
# pt : pocet bodov, v ktorych pocitame CV
{
  if (nu>0)
  {
    Xnu<-transf(X,Y,nu)$xnu
    Ynu<-transf(X,Y,nu)$ynu
  }
  else
  {
    Xnu<-X
    Ynu<-Y
  }
  koef<-optkern(nu,k,mu)
  # cvs : matica argumentov a funkcnych hodnot funkcie CV
  cvs<-matrix(0,pt-1,2)
  for (j in 1:(pt-1))
  {
    cvs[j,1]<-left+(j/pt)*(right-left)
    for (i in 1:length(Xnu))
    {
      f<-gamur(Xnu[i],X[-(i:(i+nu))],Y[-(i:(i+nu))],nu,k,mu,cvs[j,1],koef)
      cvs[j,2]<-cvs[j,2]+(Ynu[i]-f)^2
    }
    cvs[j,2]<-cvs[j,2]/(length(Xnu))
  }
  cvs
}
```

Ešte si ukážme zdrojový kód funkcie `boot()`, ktorá simuluje realizácie *bootstrap*ovej veličiny Z^*. Pre výber s opakovaním z odhadov reziduí používa funkciu `sample()`. Funkciou `boot()` bol počítaný príklad z kapitoly 2.4.

```
boot<-function(B,fn,fn0,zn0,b,x,y,steps,nu,k,mu)

# simuluje realizacie nahodnej veliciny Z* (pre maxima)
# B : pocet bootstrapovych opakovani
# fn : odhad funkcie fn v bodoch pozorovani s volbou b^{CV}_{0,k-nu}
# fn0 : odhad funkcie fn0 v bodoch pozorovani s volbou b0
# zn0 : odhad zn0
# b : vyhladzovaci parameter pre odhad radu nu,k,mu
# x,y : pozorovania
# steps : pocet bodov okolo zn0, v ktorych je ocakavana hodnota z*
# nu,k,mu : rady jadrovej funkcie

{
# spocitame odhady rezidui
  e.hat<-y-fn-mean(y-fn)
  Z.star<-matrix(0,B,2)
# okolie zn0
  xx<-zn0-steps+seq(1:(2*steps+1))-1
  koef<-optkern(nu,k,mu)
  N<-length(e.hat)
  for (i in 1:B)
  {
# vyber s opakovanim
    e.star<-sample(e.hat,N,T)
# nove bootstrapove pozorovania
    y.star<-fn0+e.star
# bootstrapovy odhad funkcie
    f.star<-gamure(x,y.star,nu,k,mu,b,xx,koef,F)
# bootstrapove maximum
    Z.star[i,1]<-xx[which.max(f.star)]
  }
  Z.star[,2]<-Z.star[,1]-zn0
  Z.star
}
```

Záver

Diplomová práca sa zaoberala problematikou odhadovania nulových bodov regresnej funkcie a jej derivácií.

Prvá kapitola pojednávala o parametrickom regresnom modeli, obsahovala najmä definíciu pseudoinverznej funkcie ako prostriedok teoreticky korektného prístupu k odhadom hodnôt závisle premennej a tvrdenie vety 1.16 o tvare konfidenčnej množiny pre nulový bod celkom všeobecnej triedy regresných funkcií.

Hlavná časť práce bola zameraná na oblasť neparametrickej regresie, ktorá je jednou z moderných štatistických metód analýzy dát. Zaoberala sa jadrovými odhadmi nulových bodov regresnej funkcie a jej derivácií v tvare, ktorý navrhli T. Gasser a H. G. Müller. Kapitoly sa venovali najmä konzistencii odhadov, ich limitným rozdeleniam, metódam hľadania optimálnych vyhladzovacích parametrov a konštrukcii intervalových odhadov nulových bodov založenej na metóde *bootstrap*.

Cieľom a účelom práce bolo najmä zhrnúť viacero známych výsledkov do kompaktnejšieho celku, hľadanie jednotiacich prvkov medzi rôznymi oblasťami venujúcimi sa danej alebo blízkej problematike, aplikácii teoretických výsledkov na reálne alebo simulované dáta.

Neparametrické odhady nulových bodov nachádzajú uplatnenie najmä v ekonometrii a bioštatistike.

Použitá literatúra

[Am85] Amemiya, T. (1985): *Advanced Econometrics.* Harvard University Press, Cambridge, MA.

[An02] Anděl, J. (2002): *Základy matematické statistiky.* Preprint. MFF UK Praha.

[AnVo92] Antoch, J., Vorlíčková, D. (1992): *Vybrané metody statistické analýzy dat.* Academia, Praha.

[Ed80] Eddy, W. F. (1980): Optimum Kernel Estimators of the Mode. *The Annals of Statistics* **8**, str. 870-882.

[Fe97] Ferreira, E., Núñez-Antón, V., Rodríguez-Póo, J. (1997): Kernel Regression Estimates of Growth Curves Using Nonstationary Correlated Errors. *Statistics & Probability Letters*, **34**, str. 413-423.

[GaMu84] Gasser, Th., Müller, H. G. (1984): Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics* **11**, 171-185.

[HaHl05] Härdle, W., Hlávka, Z. (2005): Dynamics of State Price Densities. SFB-649 "Economic Risk" Discussion Paper 2005-021. `http://sfb649.wiwi.hu-berlin.de`.

[HaMa91] Härdle, W., Marron J. (1991): Bootstrap Simultaneous Error Bars for Nonparametric Regression. *The Annals of Statistics* **19**, str. 778-796.

[JoDaPa94] Jones, M. C., Davies, S. J., Park, B. U. (1994): Versions of Kernel-Type Regression Estimators. *Journal of the Amarican Statistical Association* **89**/427, Theory and Methods, str. 825-832.

[La03] Lachout, P. (2003): *Teorie pravděpodobnosti.* Karolinum, Praha.

[Mu84a] Müller, H. G. (1984): Boundary Effects in Non-parametric Curve Estimation Models. *Compstat*, str. 84-89, Physica-Verlag, Vienna.

[Mu84b] Müller, H. G. (1984): Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes. *The Annals of Statistics* **12**/2, str. 766-774.

[Mu85] Müller, H. G. (1985): Kernel Estimators of Zeros and of Location and Size of Extrema of Regression Functions. *Scandinavian Journal of Statistics* **12**, 221-232.

[MuStSc87] Müller, H. G., Stadtmüller, U., Schmitt, T. (1987): Bandwidth Choice and Confidence Intervals for Derivatives of Noisy Data. *Biometrika*, **74**/4, str. 743-749.

[Ne95] Neumann, M. H. (1995): Automatic Bandwidth Choice and Confidence Intervals in Nonparametric Regression. *The Annals of Statistics* **23**/6, str. 1937-1959.

[OpWaYa01] Opsomer, J., Wang, Y., Yang, Y. (2001): Nonparametric Regression with Correlated Errors. *Statistical Science*, **16**/2, str. 134-153.

[Sc92] Scott, W. D. (1992): *Multivariate Density Estimation.* John Wiley & Sons, New York.

[WaGa98] Wang, W., Gasser, T. (1998): Asymptotic and Bootsrap Confidence Bounds for the Structural Average of Curves. *The Annals of Statistics*, **26**/3, str. 972-991.

[Ya03] Yatchew, A. (2003): *Semiparametric Regression for the Applied Econometrician.* Cambridge University Press, USA.

[Zv04] Zvára, K. (2004): Elektronické poznámky k prednáške k predmetu STP094 Regrese vyučovanom autorom na MFF UK v Prahe. *R & Regrese*, `http://www.karlin.mff.cuni.cz/~zvara/`[13].

[13]Jedná sa o text k prednáške, ktorý nie je určený k šíreniu. Autor tejto diplomovej práce čerpal z verzie zo dňa 15.12.2004. Text sa doplňuje, preto citácie tohoto textu nebudú obsahovať číslo strany a podobne.